How Does Scientific Progress Affect Culture? A Digital Text Analysis[*]

Michela Giorcelli, University of California – Los Angeles and NBER
Nicola Lacetera, University of Toronto and NBER
Astrid Marinoni, University of Toronto[†]

Abstract

We focus on a unique historical episode, the elaboration of the theory of evolution by Charles Darwin, to study the interplay between scientific progress and culture. We perform text analysis on a corpus of hundreds of thousands of books, with the use of techniques from machine learning. We examine, in particular, the diffusion of certain key ideas of the theory of evolution in the broader cultural discourse and imaginary. We find that some concepts in Darwin's theory, such as Evolution, Survival, Natural Selection and Competition, seldom used before, diffused in the cultural discourse immediately after the publication of *On the Origins of Species*. Other concepts such as Selection and Adaptation were already present in the cultural dialogue. Moreover, we document semantic changes for most of these concepts over time, thus providing further insights about the channels through which Darwin's theory influenced the broader discourse. Our findings provide the first large-sample, systematic quantitative evidence of the complex relation between two key factors of long-term economic growth (science and culture), and suggest that machine learning and natural language processing offer promising tools to explore this relation.

[*] We gratefully acknowledge the financial support of the National Bureau of Economic Research through the Innovation Policy Grants Program. We also thank Ryan Heuser, Graeme Hirst, Xander Manshel, Yang Xu and participants to our presentations at the NBER Productivity Lunch, the 2018 REER Conference at Georgia Tech, the Workshop in Memory of Luigi Orsenigo at Bocconi University, the University of Toronto and the 2018 Academy of Management Annual Meetings for their helpful feedback.

[†] Giorcelli: UCLA Department of Economics, 9262 Bunche Hall, 315 Portola Plaza, Los Angeles, CA 90095 (USA); mgiorcelli@econ.ucla.edu. Lacetera: University of Toronto, Institute for Management and Innovation, 3359 Mississauga Road North, Mississauga, ON (Canada); nicola.lacetera@utoronto.ca. Marinoni: University of Toronto, 105 St. George Street, Toronto, ON (Canada); astrid.marinoni@rotman.utoronto.ca.

1. Introduction

Parallel literatures in economics highlight the role of two long-term determinants of growth: scientific progress and cultural change (Alesina and Giuliano 2015, Bisin and Verdier 2011, Bush 1945, Guiso et al. 2006, Mokyr 2016, Romer 1990, Stephan 2012). To the extent that both science and culture affect economic outcomes, questions about the *relationship* between these two spheres are of economic relevance. For example, do scientific and cultural change develop independently of each other, or are they related? Are major scientific discoveries confined to debates among elites of experts, or do they percolate into popular culture and imaginary? Conversely, do different cultural environments influence scientific inquiry differently?

The investigation of whether and how scientific discoveries enter the public discourse would also address a deeper question, one about whether or not, to paraphrase Alexander Hamilton's reflections in the *Federalist Papers*, a society is based on a culture of reason and evidence, i.e. whether scientific inquiry guides social beliefs and choices.

This paper proposes the first large-sample quantitative study of the relationship between scientific progress and the broader cultural environment in which it occurs. Specifically, we analyze how the social and cultural environment of the 19th Century received one of the major scientific breakthroughs in history, the theory of evolution by Charles Darwin. We will argue that the publication of *On the Origins of Species* in 1859 was largely unplanned and therefore provides a form of natural variation, an rely on data from a large corpus of text which we analyze with natural language processing methods based on machine learning.

Casual observations reveal how intense public debates characterize scientific and technological developments; examples include the development and use of genetically modified organisms, vaccination, and the ethical concerns about the safeguard of privacy as threatened by the development of information and communication technologies. Historians and humanities scholars have long advanced the hypothesis that the social impact of scientific discoveries does not depend only on the recognition of these advances by the scientific community, but also on their broader public perception, understanding and acceptance (Bauer 2009, Cartwright and Baker 2005, Chapple 1986, Fuller 2017, Mayr 2001, Mokyr 2016, Otis 2009; Scholnick 2015).

Quantitative evidence of the interplay between scientific progress and the broader cultural climate, however, is scant. Measuring this relationship would provide social scientists information on which particular aspects and concepts that define a scientific discovery are new to the broader cultural discourse, which ones are instead already part of the overall culture, and whether scientific progress changes the nature of certain broader ideas in society. Quantitative analysis, on the other hand, presents several challenges. First, it is difficult to measure social perceptions of science and technology. Second, one would need a long-time horizon to analyze the interplay between public discourse and scientific and technological progress. Third, the plausible two-way relationship between science and culture makes it difficult to identify causal links; if, on the one hand, scientific progress can spur the diffusion and acceptance of certain ideas, on the other hand the presence and diffusion of some ideas can facilitate certain scientific discoveries.

We focus on the theory of evolution through natural selection by Charles Darwin, and examine whether the key concepts in Darwin's work emerged and diffused in the broader public discourse and social imaginary as a result of the publication of *On the Origins of Species* in 1859. Given the importance of Darwin's theory, a quantitative analysis of how his main ideas diffused in the public discourse is of relevance on its own as a source of new historical evidence and insights on how the theory of evolution affected society. In addition, the analysis of this specific episode provides a unique opportunity to explore the use of novel techniques of data collection and analysis to understand the evolution of science and culture, and to combine these new approaches with credible causal identification.

In order to estimate the effect of Darwin's theory, we exploit the fact that Darwin did not fully plan the publication date of his treatise; he had to accelerate the publication, and the public reach of his work, to keep scientific priority over it. Therefore, the timing of the diffusion of the theory of evolution was arguably exogenous. This specific context and the ensuing natural variation provide us with empirical features that are difficult to find in other cases.

Our methodology takes advantage of the development of machine learning techniques to perform digital text analysis, which we do on the Google Books corpus, a digitized collection of about eight million books. We define the publication year of *On the Origin of Species* as our reference date and concentrate our analysis on the four decades before and after it.

In Section 2, we provide a brief account of Darwin's elaboration of the theory of evolution by natural selection. We also substantiate why the publication of *On the Origin of Species* provides natural variation that allows studying the effect of Darwin's theory on the broader public discourse.

In Section 3, we describe the text-based data that we use and the techniques that we adopt to extract information about cultural evolution from these data. We first investigate whether the frequency of use of certain words and phrases changed significantly in the years following the publication of *On the Origin of Species*. We mostly consider words and expressions that, according to many accounts, represent the key concepts in Darwin's theory (Desmond and Moore 1994, Mayr 1982): Evolution, Survival, Competition, (Natural) Selection, Survival and Adaptation. The frequencies of use of these words provide a measure of the adoption and relevance of certain concepts in the public discourse. We compare, both descriptively and in a differences-in-differences econometric framework, the evolution of the frequency of use of Darwinian concepts with the frequency of a large number of words not related directly to Darwin's theory but extensively present in *On the Origins of Species*. We then analyze the semantic evolution of these words. We employ word-embedding techniques from the Natural Language Processing and Machine Learning literature.

We find two main results, which we report in Section 4. First, we some key concepts in Darwin's theory became relevant in the broader cultural discourse in the years immediately following the publication of *On the Origins of Species*: Evolution, Survival and, to a lesser extent, Competition. The patterns of diffusion of these words were similar in the non-fiction and fiction literature; this indicates a broad impact on culture as well as the social imaginary as represented, for example, by short stories and novels. Other key concepts such as Selection and Adaptation were already present in the cultural discourse. Although the relative frequency of the term Selection per se did not vary around the publication of *On the Origins of Species*, the expression Natural Selection was virtually nonexistent in the literature before 1859 and diffused rapidly thereafter; this suggests a potential change in the way the term Selection was used.

The second set of results, in fact, concerns semantic changes. Of interest is the increase in semantic association between words such as Competition (or Struggle) and Life, as well as between Life and Adaptation, immediately following the publication of *On the Origins of Species*. This is consistent with Darwin's theories affecting the perception of what existence means and how it unfolds. Furthermore, the term Adaptation became, over the 19th Century, less related to physical terms (such as Mechanism) and increasingly related to concept related to living beings (such as Organism and Reproduction). The term Evolution, which came mostly from chemistry and physics in the first half of the 1800s, later in the century related more to concepts from biology as well as social and human subjects, indicating a broader reach of this idea in society. Finally, Selection became more similar in meaning to other "Darwinian" words, such as Survival, Variation, Fittest and Heredity.

At least in this case, therefore, scientific progress lead to the diffusion of some concepts in the broader cultural discourse, as represented by a very large corpus literary production that we were able to analyze, and also affected the use and meaning of concepts that were already part of a culture. To the extent that a culture that values scientific inquiry and evidence is more likely to promote economic development (Mokyr 2016), a channel through which this appreciation may occur is precisely through scientific progress.

In Section 5, we provide concluding remarks and direction for future research.

Related literature. The stream of literature that is closest to our work includes studies of how different cultures are more or less open to scientific and technological change, and how certain scientists may introduce new sets of beliefs in a population. Mokyr (2013, 2016), in particular, defines "cultural entrepreneurs" those scientists who put in motion broader cultural changes. Our paper provides an empirical approach to study this form of cultural entrepreneurship.

We also contribute to the growing use of "text as data" in economics, which is developing especially in such fields as finance, marketing, political economy and the study of media (Gentzkow et al. 2018, Jelveh et al. 2014). Economists of science, productivity and innovation have recently begun to rely on these sources of information and related techniques (Balsmeier et al. 2018; Bandiera et al. 2017; Catalini et al. 2015; Kelly et al. 2017).

Scholars in linguistics and literary criticism are also increasingly employing computerized text analysis to answer questions about the evolution of literary genres and styles, and semantic changes of words and concepts. Instead of relying on the direct reading of an inevitably limited set of texts from which to offer general insights and interpretations, this line of research is based on the automated or "distant" reading of a much larger set of digitized texts (Heuser and Le-Khac 2011, Heuser 2016, Moretti 2013, Wilkens, 2015). Cohen (1999) uses the expression "great unread" to indicate the large quantity of books and texts that normally scholars do not study, but that, as a whole, represent the broader social and cultural climate at a given time. In addition to literary analysis, an area of study known as "cultural analytics" or "culturomics" also explores the evolution of culture through text analysis (Aiden and Michel 2014, Manovich 2009, Michel et al. 2011), and in particular through the study of changes in the frequency and meaning of certain words and expressions over time. To our knowledge, there are no applications of these approaches to studying the public perception of science.

Finally, our work also relates to the literature on the role of institutions in the diffusion of ideas and innovation (Abramitzky and Sin 2014). Our paper looks at the impact of scientific advancements on the perception of key ideas and concepts in society, and on how these ideas and concepts were already permeating the public discourse.

2. Historical Background and Identification

> *"It is doubtful if any single book, except the 'Principia,' ever worked so great and so rapid a revolution in science, or made so deep an impression on the general mind."*
> Obituary for Charles Darwin, Proceedings of the Royal Society of London, 1888.

2.1. The Development of Darwin's Theory of Evolution

Charles Darwin's interest in the evolution of living organisms largely developed during his voyage, from 1831 to 1836, on the HMS Beagle, a ship of the Royal Navy. Over those five years, Darwin collected fossils from the places that he visited and observed their geographical distribution. He was particularly interested in the geographical distribution of wildlife and fossils that he collected in the voyage. Although his early elaborations built on previous theories (such as Lamarck's and Chambers') and considered the possibility of the transformation of one species into another (transmutation), he then developed his own theory of evolution based on the natural selection of the most adaptive (innate) characteristics of a species. Small, gradual variations within a species would emerge randomly, and would eventually lead to branching of new species. Competition for resources and adaptive capacities would determine whether and where a particular species would be more likely to thrive. The developments in genetic research since the mid-20th Century provided corroboration and foundations to Darwin's evolutionary theory. (Desmond and Moore 1994, Mayr 1982).

In addition to being one of the greatest scientific breakthroughs in history, there is a perception that Darwin's theory of evolution had a wider cultural reach (Desmond and Moore 1994, Mayr [1982, 2001], Fuller 2017,). In particular, the ideas of competition for resources, common origins of species, and random variation implied the absence of a teleology or (benevolent) design, that is, a very different conception of nature and of God.[3]

The likely common origins of all species, moreover, eliminated any idea of superiority of humans as compared to other living beings, and, within humans, of a race with respect of another. Fuller (2017), for example, argues that Darwin's theory had a major influence on the debate over race, slavery and discrimination in the United States, thus hinting at a major role of this scientific breakthrough in the evolution of American society. Mokyr (2013, 2016) includes Darwin among a small set of "cultural entrepreneurs", i.e. scientists whose discoveries questioned deeply held cultural and popular beliefs.

These accounts, however, focus on a narrow set of literary contributions or on cultural debates mostly restricted to a scientific, political and economic elite; this makes it hard to advance inferences about the broader cultural impact of this scientific advance, and about the cultural climate that preceded that breakthrough. Our approach to answering these questions relies on a massive corpus of literary work (fiction and non-fiction), and therefore offers a methodological contribution that allows going beyond the analysis of a small set of texts and authors as a way to extrapolate general cultural views and trajectories.

2.2. Identification Strategy

Some features of how Darwin made his work public enhance our ability to identify the impact of Darwin's work on the broader cultural discourse. Although Darwin developed his theory over a long period, there is a precise time at which Darwin's theory reached a broader public, and this is 1859, the year of publication of *On the Origin of Species*.[4] This publication date was largely unplanned. Darwin proceeded slowly initially and had to deal with sickness and deaths in his

[3] Research in literary criticism analyzed how the production of certain poets and novelists, began to reflect ideas of a different role that nature had in its relationship with humans and the environment. Similarly, studies of the literary production prior to the publication of *On the Origin of Species* point out how some of Darwin's ideas connected to images already developed by these writers. A frequently cited example is the work of Alfred Tennyson, and in particular his poem *In Memoriam*, published in 1850. Scholars also investigated the connections between broader worldviews, such as Enlightenment and Romanticism, on Darwin's ideas (Cartwright and Baker 2005; Chapple 1986; Gianquitto and Fisher 2014; Lansley 2016; Otis 2009; Richards 2013; Scholnick 2015).

[4] The year 1859 saw also the publication of other important works, John Stuart Mill's *On Liberty*, Tennyson's *Idylls of the King*, Eliot's *Adam Bede* and Dickens' *A Tale of Two Cities*. These publications make it harder to identify a connection between the publication of *The Origins of Species* and changes in the public discourse. However, in our study, we focus on rather specific concepts that are central in Darwin's work but not in the other works mentioned above; we also consider the presence of those concepts in the public discourse *before* 1859.

family that delayed him. However, eventually he "rushed" in order not to lose priority over Alfred Russel Wallace, who was researching on the same topics and had sent Darwin some of his writings that used similar concepts and reached similar conclusions about natural selection as the theory that Darwin elaborated.

The book and Darwin's overall theory received almost immediate attention and fast diffusion, also thanks to presentations at prestigious scientific meetings such as the Linnaean Society (of a joint paper with Wallace in 1858) and the British Association for the Advancement of Science (in 1860), as well as reviews in the popular press (see for example Gray 1860, Huxley 1859).

The unplanned publication date of Darwin's theory provides the main source of exogenous variation for our empirical study. The rapid diffusion of the theory gives us an opportunity to observe variation in the diffusion of the main concepts, and to establish which ones were especially novel and had an independent impact on the broader public discourse.

To be sure, *On the Origins of Species* was not the first treatment of evolution. Darwin's theory was novel in several ways and more coherent than previous ones, but earlier in the 19th Century some related ideas were already "in the air" – examples include the work of Lamarck, the anonymous *Vestiges of the Natural History of Creation* (later attributed to the Scottish journalist Robert Chambers), and of course the work of Alfred R. Wallace. Our empirical strategy, however, allows assessing whether the publication of Darwin's book represented a discontinuous change in the cultural discourse, or whether some of the main concepts, perhaps through the work of some of his predecessors, were already embedded in the public discourse.[5]

3. Data and Methods

To examine the diffusion and semantic evolution of scientific concepts over time, we exploit the increasing availability of digitized historical text corpora, as well as new tools of natural language analysis. Our first step is to compute relative frequencies of some key words that embody the main concepts advanced in Darwin's theory of evolution, and that in fact Darwin used extensively in his own work. These frequencies represent a basic measure of the adoption of certain ideas in the broader cultural and social discourse. The second step of our investigation focuses on word embeddings, which are widely used in the Natural Language Processing and Machine Learning

[5] See in particular Desmond and Moore (1994) for details on the personal and intellectual biography of Darwin.

literature as an effective tool for the analysis of semantic change. Several studies, especially in computational linguistic, computer science, and digital humanities, have validated these ways to measure cultural change (Aiden and Michel 2014, Manovich 2009, Michel et al. 2011, Roth 2014).

Word Frequencies. We rely on Google N-Grams[6] (Lin et al. 2012) to assess how frequencies of words changed over time. The Google N-Grams data is the result of the Google Book project whose aim is to build a vast collection of digitized books in partnership with major libraries[7]. First released in 2010, the data consists of a set of corpora of roughly eight million books, an estimated 6% of all books ever published (Lin et al. 2012). The texts cover roughly a 500-year span and they are continuously updated. The Google Books data includes different corpora and languages (besides English: Italian, French, German, Spanish, Russian, Hebrew, and Chinese). The English corpus alone has half a trillion words in it. The data include both fiction and non-fiction books, but not periodicals, and is aggregated depending on the number of terms considered; for instance, the 1-ngram dataset includes single words and their frequency in a given corpus, and n-grams include combinations of n words and their frequency. We compute frequencies from 1-ngram and 2-ngrams data for each year and express them in per-million-words terms.

The ability to separate fiction and non-fiction literature is particularly relevant to us for two main reasons. First, one critique to the N-gram (and Google Books) corpus is that it may over-represent scientific texts (Pechenick et al. 2015). In our study, an uptake in the frequency of words related to Darwin's theory may just reflect a disproportionate increase over time of the corpus of scientific books (included in the non-fiction category). Second, separating fiction and non-fiction literature enables the analysis of different types of relationships between Darwinian science and broader culture. The use of Darwinian concepts in the non-fictional literature may better represent higher-educated or more erudite conversations. Conversely, given the diffusion of the novel and the relatively high literacy rates especially in England and the United States in the 19th Century, fictional literature may better measure ideas in the broader social imaginary (Armstrong 1987, Winans 1975).

[6] Available at: http://books.google.com/ngrams.

[7] http://books.google.com/googlebooks/library/partners.html

Semantic evolution and word embeddings. The analysis of word frequencies is informative, but it does not provide insights about associations between words over time. To this aim, we employ a distributional natural-language processing technique, known as word embeddings, which is able to capture semantic and contextual changes of words in a given period. Key to word embeddings is the representation of words as vectors, with the values in a vector reflecting the co-occurrence of the focal word with other terms. Consider a vocabulary with *V* distinct words in it. Represent each word *w* as a V-dimensional vector where each entry represents a measure of how likely each other word is to occur within a window of *m* words around *w*.[8] Machine-learning algorithms allow predicting the words surrounding *w* (or context words). We adopt, in particular, a Word2Vec approach (SkipGram with negative sampling; Mikolov et al. 2013). The model on which we rely computes estimates of parameters θ that solve:

$$\arg\max_{\theta} \prod_{w \in T} \prod_{c \in C(w)} p(c|w;\theta). \qquad (1)$$

The term *w* indicates a focal word in a corpus, and *c* represents a context word included in *C(w),* the set of possible context words, i.e. the words that appear within a window of *m* words around *w*. The parameters of the models are set such that the probability of context words appearing near the target words is as high as possible. After expressing equation (1) as a negative log-likelihood, we parametrize the model following the neural-network literature, using a soft-max function:

$$\mathrm{p}(c|w;\theta) = \frac{e^{v_c\, v_w}}{\sum_{c' \in C} e^{v'_c\, v_w}}, \qquad (2)$$

where v_c and v_w are vector representation of *c* and *w* respectively. *C* is the set of all possible contexts. The training process starts with random vectors that are "pulled closer or apart" depending on the actual word co-occurrence. The final vectors satisfy some "linearity" features in the relationship between, for example, the singular and plural form of a word, or feminine and masculine versions. Using a frequent example in the literature, we expect, when the words *king, kings, queen, queens, man* and *woman* are in distributed vector form, that the following holds: *(king – kings) ≈ (queen – queens)* and *(king – man) ≈ (queen – woman)*.

The more "similar" two word vectors are, the closer the semantic association of the two words.[9] The main metric to compare the vector representations of words is the cosine distance

[8] Smaller windows tend to capture functionally similar words (e.g., 'respect' and 'deference'), whereas larger windows capture context relatedness or topic similarity (e.g., 'respect' and 'love') (Levy and Goldberg, 2014).

[9] Embeddings can measure close semantic relationships between words as well as more global ones. For instance, beyond successfully measuring shifts in word meanings over time (Hamilton et al. 2016), embedded vectors have also

(Dubossarsky et al. 2015; Gulordava and Baroni 2011; Jatowt and Duh 2014; Kim et al. 2014; Kulkarni et al. 2015). Call γ the angle (generalized to *N*-dimensions) between two *N*-dimensional vectors $u = (u_1, ... u_N)$ and $v = (v_1, ... v_N)$. Then, $u'v = \sqrt{\sum_{i=1}^{N} u_i^2} * \sqrt{\sum_{i=1}^{N} v_i^2} * \cos(\gamma) = \|u\|\|v\| \cos(\gamma)$, or: $\cos(\gamma) = \frac{u'v}{\|u\|\|v\|} \in [-1,1]$. The more similar the two vectors, the closer to one the cosine. We use previously trained Word2Vec embeddings resulting from the n-grams distributed by Google Books (Hamilton et al., 2016). Figures are available for every decade between 1800 and 1990 and data are specifically designed to enable comparisons across decades. We use a context window of four (on each side), and set the parameters as suggested by Levy et al. (2015)[10]. In general, these choices for the window and the parameters are expected to measure semantic changes and more generally cultural shifts (Hamilton et al., 2016).

4. Findings

We first describe the evolution of the relative frequency of certain selected words and two-word expressions as measures of the diffusion of certain concepts in the public discourse around the time of the publication of *On the Origin of Species* in 1859. The second part of the analysis focuses on semantic evolution.

4.1 Word Frequency and Diffusion of Concepts

4.1.1 Darwinian and "Control" Concepts

Graphical Analysis. We consider terms (1-grams) that, from many accounts (Desmond and Moore 1994, and Mayr 1995), as well as our own reading, represent the key concepts in Darwin's theory: Evolution, Selection, Adaptation, Competition, Survival, and the expression (2-gram) Natural Selection. Figure 1 reports their frequency of use, per million words, in each year between 1820 and 1899, separately in fiction and non-fiction books.[11] We scale the y-axes differently for the two categories in each graph.

been used to track demographic and occupational social shifts (e.g., Garg et al. 2017) and gender stereotypes (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017).

[10] See also Hamilton et al. (2016) for a discussion on the pre-processing methods and parameters.

[11] We initially included also the word Mutation, but then opted to discard it its occurrence was too low throughout the period of interest to allow for meaningful analyses

The expression Natural Selection, perhaps the most defining of Darwin's concept, was virtually non-existent in both the fiction and non-fiction literature before 1859 and experienced a significant increase in use since then. On the one hand, this may not be surprising, precisely because of the close association of Darwin with the idea of natural selection. On the other hand, we may consider the significant increase in the diffusion of this concept immediately after the publication of *On the Origin of Species* as a validation of our approach; this initial analysis of frequencies seems to capture what we might have expected.

Moving to other Darwinian concepts (perhaps not as tightly associate to Darwin as Natural Selection), also Evolution and Survival entered the public discourse in the years immediately following the publication of *On the Origin of Species*. The concepts that underlie these words and expressions, therefore, generated interest not only in specialized or more educated circles, but plausibly also in the more popular cultural context. Interestingly, the diffusion of these concepts in the fiction literature seems to have lagged the diffusion in the non-fiction literature by a few years. Competition was already present in the first part of the 19th Century, but especially in the non-fiction literature, and experienced an increase in frequency after about 1860.

Selection and Adaptation, in contrast, did not see a further increase in relative frequency around the publication of *On the Origins of Species*; Adaptation reached a stable relative frequency in the 1840s, whereas the relative frequency of Selection was constantly increasing since the early 19th Century. Note how Selection was already increasing its presence in the cultural discourse before 1859, whereas Natural Selection appeared after the publication of *On the Origins of Species*. This suggests the possibility that, after 1859, the word Selection might have experienced semantic changes, i.e. a change of meaning in the public discourse. We will explore this below.

In Figure 2, we display the relative occurrence of some terms that of frequent use in general and in the sciences, are not specific to Darwin's theory, and appear very frequently in *On the Origins of Species*. In looking at these terms, our objective is to assess whether there were general trends in the use or diffusion of scientific concepts. The words that we consider in the figure are Number, Life, Animals, Flowers, Plants and Nature. For none of these words was there any discernible change in diffusion in the decades immediately preceding and following the publication of *On the Origin of Species*. These "generic" words are a subsample of the 100 nouns whose frequency we use as counterfactuals in the regression analyses that we describe below.

Regression Results: Word-by-Word Time Series. Table 1 reports estimates from regressions of the yearly relative frequency of use of each of the Darwinian words and phrases, as well as of the subsample of six generic words that we represented in the graphs above. We rely on the following models, where the outcome variable is the frequency per million words of each word *w* in year *t*, expressed either in absolute terms (model 3 below and panel A of Table 1) or in natural log transformation (model 4, panel B; we added 0.01 to each value of the frequencies):

$$y_{wt} = \alpha_w + \beta_w t + \gamma_w \mathbf{1}(t > 59) + \delta_w \mathbf{1}(t > 59) * (\mathrm{t} - 59) + \varepsilon_{wt}; \tag{3}$$

$$\ln(y_{wt}) = \alpha_w + \beta_w \ln(t) + \gamma_w \mathbf{1}(t > 59) + \delta_w \mathbf{1}(t > 59) * (\ln(\mathrm{t}) - \ln(59)) + \varepsilon_{wt}. \tag{4}$$

We define the time trend *t* as the current year *minus* 1800, therefore it takes values 20, 21,…, 99. The coefficients γ_w and δ_w measure, respectively, "step" and slope changes in frequency before and after 1859.[12] In the log-transformed model, these changes are in relative terms, and therefore easier to compare across words.

The estimates, both in absolute and relative terms, reinforce the visual evidence from Figures 1 and 2. For the "Darwinian" terms discussed above, the increase in slope after 1859 is significant and especially large for Natural Selection, Evolution, Survival and Competition. We do not detect any specific pattern related to the publication of *On the Origins of Species* for the six "control" words. We ran the same regressions as in model 4 for all of the 100 most frequent nouns in *On the Origins of Species*. More precisely, one of these 100 nouns is Selection; therefore, the actual number of control words is 99 (the list is in Appendix Table A1). Figure A1 in the appendix reports the estimates of γ_w and δ_w for each of these nouns. The vast majority of the estimates is not statistically significant, and the estimates that are significant are split between negative and positive, showing, again, no detectable pattern. We will rely on the full set of nouns also in the differences-in-differences analyses below.

Table 2 reports regression estimates of the following model:

$$\begin{aligned}\ln(y_{wt}) = {} & \alpha_w + \beta_w \ln(t) \quad + \gamma_w \mathbf{1}(t > 59) + \delta_w \mathbf{1}(t > 59) * (\ln(\mathrm{t}) - \ln(59)) + \\ & \alpha_{wf} \mathbf{1}(fiction) + \beta_{wf} \ln(t) * \mathbf{1}(fiction) + \gamma_{wf} \mathbf{1}(t > 59) * \mathbf{1}(fiction) + \\ & \delta_{wf} \mathbf{1}(t > 59) * (\ln(\mathrm{t}) - \ln(59)) * \mathbf{1}(fiction) + \varepsilon_{wt}\end{aligned} \tag{5}$$

[12] In the Appendix, we estimate a model where we allow for step changes and changes in slope for every decade between 1820 and 1899.

We run this regression for each word on *N=160* observations, two per each year with one pertaining to non-fiction books ($\mathbf{1}(fiction) = 0$), and one reporting relative frequencies of the focal word in fiction books ($\mathbf{1}(fiction) = 1$). Especially for the words and expressions that experienced an increase in diffusion immediately after 1859, the regressions provide support to the observations that we made with respect to Figure 1 above: the diffusion concerned both the non-fiction and fiction literature.[13,14]

Regression Results: Differences in Differences. After having studied the diffusion over time of each word separately, we proceed with some differences-in-differences analyses where we estimate the aggregate diffusion patterns of Darwinian and generic scientific concepts before and after 1859. We perform these analyses in two ways.

First, in Table 3 we report the estimates from regressions where, for each year, we sum up the frequencies of occurrence (in a given year) of the six Darwinian concepts on the one hand, and of the 99 "control" nouns on the other hand, and compare the trend in the aggregate diffusion before and after 1859. Because the aggregate frequency of the generic words is much higher than the frequency of the Darwinian concepts pooled together, to make more immediate comparisons we transform these frequencies into their natural logarithms and include the logarithm of the time trend in the regression analyses. Therefore, we compare scale-free elasticities. In this analysis, we also pool together fiction and non-fiction books. The regression model that we estimate is as follows:

$$\begin{aligned} &\ln{(y_{wt})} = \alpha_w + \beta_w \ln(t) + \gamma_w(\ln(\mathrm{t}) - \ln(59)) * \mathbf{1}(t > 59) + \\ &\delta_w \mathbf{1}(Darwinian\,word) + \theta_w \ln(t) * \mathbf{1}(Darwinian\,word) + \end{aligned} \qquad (6)$$

[13] Appendix Table A3, where we report estimates with specific slopes and steps for each decade between 1820 and 1899 instead of only one "cut" in 1859, at every decade, also confirms the delay in diffusion in the fiction literature that we observed in the graphical representations above.

[14] We also explored the evolution of the concept of gradualism as applied to the type of change and evolution that Darwin considered. The idea of small, continual changes at the basis of the evolution for species and more generally of biology is key in Darwin's work; it is also a philosophical contribution or worldview. We also considered expressions such as Gradual Change, Gradual Adaption, Gradual Divergence and Gradual Mutation. The very low relative frequencies of these expressions in our corpus, however, do not allow making any clear inference. Data are available upon request. There was no strong trend throughout the 19th Century, especially for the most frequent of the di-grams (Gradual Change), nor any specific change in adoption rates around 1860. This lack of a clear effect may be consistent with the idea that this concept was already part of a "Victorian" view of society and this contributed to the acceptance of several aspects of Darwin's theories. But, again, given the very low overall frequencies, we need to be cautious in drawing conclusions.

$\lambda_w(\ln(\mathrm{t}) - \ln(59)) * \mathbf{1}(t > 59) + \mu_w(\ln(\mathrm{t}) - \ln(59)) *$
$\mathbf{1}(Darwinian\ word) * \mathbf{1}(t > 59) + \varepsilon_{wt}.$

The data thus include *N=160* observations, two for each year, with one reporting information about the generic words ($\mathbf{1}(Darwinian\ word) = 0$), and the other about the six Darwinian concepts ($\mathbf{1}(Darwinian\ word) = 1$). Columns 1 and 2 of Table 3 display estimates of a simplified version of the model, were the left-hand-side variable is the natural logarithm of the sum of frequencies of Darwinian and generic terms separately, regressed on a time trend and the interaction between the indicator for years greater than 1859 and the difference between the current year and 1859. Estimates of the parameters of the full model 6 are in column 3. The estimate on the coefficient on the interaction between the indicator for Darwinian words, the indicator for the post-1859 period and the difference between the current year and 1859 (μ_w) is positive, large and statistically significant, indicating a much larger relative increase in the frequency of Darwinian concepts after 1859. The estimate of the parameter θ_w is significantly smaller than the estimate of μ_w, but it is also statistically different from zero; this indicates that also before 1859, the frequency of Darwinian concepts was increasing at a higher rate that the combined 100 generic terms. This is likely due to the trend and diffusion that some Darwinian terms, such as Selection and Adaptation, were experiencing also in the first half of the 19th Century.[15] The trend, however, clearly experienced and additional, fast acceleration after the publication of *On the Origins of Species*.

Second, we consider a model where the outcome variable is the annual frequency (from 1820 to 1899) of the six Darwinian concepts and the 99 control nouns separately (*N=8,400*), and we estimate the average difference in frequency for the Darwinian words and the generic words per each decade:

$$\ln(y_{wt}) = \alpha_w + \beta_w \mathbf{1}(Darwinian\ word) + \textstyle\sum_{j=2}^{4} \gamma_i \mathbf{1}(j0 \le t \le j9) +$$
$$\textstyle\sum_{j=6}^{9} \gamma_j \mathbf{1}(j0 \le t \le j9) + \sum_{i=2}^{4} \delta_i \mathbf{1}(j0 \le t \le j9) * \mathbf{1}(Darwinian\ word) + \qquad (7)$$
$$\textstyle\sum_{j=6}^{9} \delta_j \mathbf{1}(j0 \le t \le j9) * \mathbf{1}(Darwinian\ word) + \varepsilon_{wt}$$

Figure 3 displays the estimates of the δ_j coefficients and 95% confidence intervals. The omitted time category is the decade 1850-59 ($50 \le t \le 59$). This analysis provides further evidence of the

[15] If, for example, we exclude Adaptation and Selection from computing the aggregate frequency of the Darwinian concept, the estimate of θ_w declines from 0.36 to 0.08, whereas the estimate of μ_w increases from 1.74 to 3.08.

different patterns of diffusion of the Darwinian words immediately following 1859, compared to statistically insignificant differences before the publication of *On the Origins of Species*.

4.1.2 Lamarck and Darwin; Transmutation and Evolution

If a word frequency analysis is a valid way to measure the diffusion and acceptance of an underlying scientific theory in the broader cultural discourse, then this analysis should also be able to identify the decline of certain theories. A natural comparison to Darwin's elaboration is Lamarck's theory of the transmission of acquired traits. We plot the relative frequency of the use of the words "Darwin" and Lamarck" in books. Because Lamarck was French (and was writing in that language), we do the same exercise also on the corpus of French books. For English texts, we further isolate the frequency in the fiction literature; this is not possible for texts in French in Google Ngrams. Figure 4 reports the frequency graphs. The frequency of the word Darwin became increasingly greater than the frequency of Lamarck both in the English and French literature; in the latter case, the frequency of Darwin surpassed that of Lamarck soon after 1859. Darwin seems to have had a larger presence in the English fiction literature than Lamarck, too. We also compare in Figure 4 two terms that related to the study of the emergence and development of species: Evolution and Transmutation. Although Evolution, which we already analyzed above, is typically associated with Darwin's work, earlier works in biology (including some of Darwin's) used the term Transmutation to characterize (gradual or discrete) transformations of plants and animals. By comparing these two words, we want to assess whether the broader literature and cultural discourse also picked up the "newer" word to express these changes. For books in French, we consider the word Transformism (Transformisme in French), which was used by Lamarck. The general pattern is that Evolution became progressively more frequent than Transmutation, with a significant change in frequency after 1859. Transmutation ad Transformism were very rarely used both before and after 1859; therefore, this comparison is less informative, overall, than the one between the frequency of us of the words Darwin and Lamarck.

4.2 Semantic Changes

Looking at frequency of use as a measure of the interplay between a major scientific discovery and the broader cultural climate is a natural first step for our analysis. However, the role of a particular construct does not only depend on how often that construct occurs in books. Words can change their meaning over time. These changes, even keeping frequency constant, provide further evidence of cultural evolution potentially linked to certain scientific events.

Figure 5 introduces the second part of our study, where we move from the analysis of the frequency of use of certain words, expressions and the concepts underlying them, to the analysis of whether the semantic evolution of certain words and concepts, to see whether this evolution occurred in ways that we can relate to the elaboration of Darwin's theory. In the graphs, the horizontal axis reports decades (the time unit of reference as describes in Section 3), and the vertical axis indicates the cosine between the two-word vectors of interest.

One aspect of Darwin's theory is that life (or existence) includes adaptation, as well as competition, among its defining aspects. We do see an increase in the semantic association between Life on the one hand, and Adaptation, Struggle and Competition on the other hand, especially after 1859. For Life and Struggle we see a trend since the early 19th Century. Several of the studies mentioned above that relate Darwin's work to the Romantic literary climate of the first half of the 19th Century, with a more tumultuous view of nature in particular, seem therefore to have captured a more general trend. Greater cosine similarity between Survival and Competition started in the 1860s and increased since then. Finally, a controversial implication of Darwin's theory is that evolution applies to humans in the same way as it applies to other animals; although Darwin did not explicitly treat the human species in his 1859 book, this was the topic of his 1871 *The Descent of Man and Selection in Relation to Sex*. The semantic evolution of the word Human shows an increase in its similarity with Animal especially in the late 1800s.

A second analysis of semantic changes focuses on some of the key words and concepts that we considered so far. Instead of investigating the similarity of these words with a select sample of other concepts, we "let the data speak" by determining, for each decade, the words with the highest semantic connection (cosine similarity) to these key words. Figures 6 through 11 report the findings for the words Adaptation, Competition, Evolution, Nature, Selection and Survival. We excluded from the rankings of semantic similarity the words that had the same root as the focal key word as well as the most obvious synonyms (e.g. Compete or Competitor for Competition); we also defined a lower bound to the relevant cosine similarity to be equal to 0.05. The closer a word is to the horizontal (time) axis in the figures, the closer to one the cosine similarity. Finally, we use a "color system" to classify words according to some broad category; in addition to being interest in changes in the type of most similar words, we also want to assess whether, for example, concepts more distant from Darwin disciplines related to Darwin’s major concepts and whether these similarities evolved over time.

The figures identify a few interesting facts. First, the term Adaptation became, over the 19th Century, less related to physical or “mechanical” terms (such as Mechanism) and increasingly similar to concepts that represented living beings (such as Organism and Reproduction).

Second, the biggest changes in meaning and association concern the word Evolution. In the first half of the 19th Century, the terms that were closest to Evolution came mostly from chemistry and physics. Later in the 1800s concepts from biology as well as related to human society were semantically more similar to Evolution. Examples include Social and Progress. Note also how the word Darwinian itself became closely associated with Evolution; this is consistent with a direct role of Darwin’s theory in changing the meaning of this concept.

Third, Selection appeared more closely related to the concept of Choice (and qualification for the choice such as “careful” or judicious”) in the first half of 1800; the similarity in meaning with Choice remained also later, but in the broader literature, Selection became more similar in meaning to other specific “Darwinian” words, such as Survival, Variation, Fittest and Heredity.

Fourth, very few words had a similarity in meaning with Survival, likely because the word itself was only rarely used in the first half of the 19th Century. Later in the century, the word was increasingly associated in the overall literature to other concepts related to evolutionary theory, notably Fittest, Evolution, Struggle and Selection. The increasing relatedness with Fittest toward the end of the 1880s is likely due also to the publication of the *Principles of Biology* by Herbert Spencer in 1864, where this concept applies also to society and ethics and not only to the natural sphere. Competition, in contrast, maintained an association with a stable set of words, mostly related to production and markets, throughout the century.

Finally, Nature is perhaps too generic (and was more widely used) of a term to expect a close relation with specific concepts. Interestingly, however, words such as Divine and Perfection disappear from the concepts most closely related to Nature in the second half of 1880.

5. Conclusions

To the extent that both cultural and scientific change are major drivers of long-term economic development, the investigation of how these two phenomena coevolve promises to offer a deeper understanding of their role in enhancing growth.

We focused on one specific scientific breakthrough, the theory of evolution via natural selection of Charles Darwin, and explored its impact on the public discourse in society. There is a diffused perception that Darwin's theory affected culture in many different ways, from affecting the interpretation of the role of nature to influencing ideas about race and equality among humans. Existing accounts, however, largely rest on qualitative evidence of debates among scientists or elites in society, whereas little is known about the diffusion of Darwin's ideas into society at large. Arguably, to affect cultural change (to be, in the terminology of Mokyr [2013, 2016] a cultural entrepreneur), a scientist should have an impact on the imaginary of a broader population. Moreover, it is difficult to identify, from existing accounts, which Darwinian concepts were actually novel in the cultural discourse, and which ones were already part of it. We address these challenges by analyzing the diffusion and the semantic evolution of the key words and phrases that embody Darwin's main concept in hundreds of thousands of books, with the use of techniques from machine learning. We rely on the largely unplanned publication date of *On the Origin of Species* as source of natural variation, and compare the use of these words and phrases with more generic terms that Darwin used.

Our analysis shows that the key concepts expressed by Evolution, Survival, and Natural Selection diffused in fiction and non-fiction literature immediately after the publication of *On the Origins of Species*. Competition, a theme already present in the broader literature, diffused significantly more rapidly after 1859. Other key concepts such as Selection and Adaptation were already gaining relevance in the cultural discourse before 1859. The adoption of some of these words and phrases in the broader cultural conversation led also to a change in the meaning of the concepts, providing further evidence of the impact of Darwin's theory in society at large.

Our approach has several inductive and descriptive aspects. Although the choice of the concepts on which to focus may seem somewhat arbitrary, we based our selection on the main topics that Darwin developed in his treatise, as well as on the analysis of several interpretations of Darwin's theory of evolution. Moreover, it is generally hard to provide causal identification with this type of analysis. However, the unplanned publication date of *On the Origins of Species,* the reliance on very large amount of data, and the consistency in the patterns of different words, phrases and concepts, give us some confidence about the nature of the patterns that we established.

Finally, this is a single case study, and generalizations about the relationship between major scientific discoveries and their cultural reception are difficult to make. Empirical approaches enabled by machine learning techniques provide promising tools to explore this relationship beyond the specific historical episode on which we focus. In addition to making it possible to analyze a vast amount to textual data and to relate them to specific underlying ideas, these approaches allow identifying, for example, which concept of a novel scientific contribution had influence beyond the specific scientific domain, and whether and how a scientific breakthrough changes the perception of certain ideas in society. Examples of relevant scientific breakthroughs include the theory of relativity or the indeterminacy principle in physics, the discovery of the DNA, and the emergence of biotechnology and genetic engineering. In fact, one could go beyond scientific discoveries and employ a similar approach to explore the cultural antecedents and effects of new technologies as well as of new industries, such as computers and the Internet (see for example Turner 2010).

References

Abramitzky, R. and Sin, I. (2014). Book translations as idea flows: The effects of the collapse of communism on the diffusion of knowledge. *Journal of the European Economic Association*, 12(6):1453-1520.

Aiden, E. and Michel, J.B. (2014). *Uncharted: Big data as a lens on human culture*. Penguin.

Alesina, A., and Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, *53*(4), 898-944.

Armstrong, N. (1987). *Desire and domestic fiction: A political history of the novel*. Oxford University Press.

Balsmeier, B., Li, G.C., Assaf, M., Chesebro, T., Zang, G., Fierro, G., Johnson, K., Lück, S., O'Reagan, D., Yeh, B. and Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economics & Management Strategy*, forthcoming.

Bandiera, O., Hansen, S., Prat, A., & Sadun, R. (2017). CEO Behavior and Firm Performance (No. w23248). National Bureau of Economic Research.

Bauer, M. W. (2009). The evolution of public understanding of science-discourse and comparative evidence. *Science, technology and society*, 14(2):221-240.

Bisin, A., & Verdier, T. (2011). The economics of cultural transmission and socialization. In *Handbook of social economics* (Vol. 1, pp. 339-416). North-Holland.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, pages 4349-4357.

Bush, V. (1945). *Science, the endless frontier: A report to the President*. US Govt. print.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183-186.

Cartwright, J. H. and Baker, B. (2005). Literature and science: Social impact and interaction. *Abc-Clio.*

Catalini, C., Lacetera, N., and Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences*, 112(45):13823-13826.

Chapple, J. (1986). *Science and Literature in the 19th Century*. London: Macmillan.

Cohen, M. (2002). *The sentimental education of the novel*. Princeton University Press.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493-2537.

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4):447-464.

Desmond, A. J., & Moore, J. (1994). *Darwin*. WW Norton & Company.

Dubossarsky, H., Tsvetkov, Y., Dyer, C., and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. *NetWordS*, pages 66-70.

Fuller, R. (2017). *The Book that Changed America: How Darwin's Theory of Evolution Ignited a Nation*. Penguin.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2017). Word embeddings quantify 100 years of gender and ethnic stereotypes. *arXiv preprint* arXiv:1711.08412.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2018). Text as data. *Journal of Economic Literature*, forthcoming.

Gianquitto, T., & Fisher, L. (Eds.). (2014). *America's Darwin: Darwinian Theory and US Literary Culture*. University of Georgia Press.

Gopnik, A. (2010). *Angels and ages: A short book about Darwin, Lincoln, and modern life*. Vintage.

Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of political economy*, *102*(5), 912-950.

Gramsci, A. (1948). 2003. Selections from the prison notebooks. *The civil society reader. Hanover and London: University Press of New England*.

Gray, A. (1860). Darwin on the Origin of Species. *The Atlantic*, July issue.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67-71. Association for Computational Linguistics.

Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic perspectives*, *20*(2), 23-48.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint* arXiv:1605.09096.

Harrison, L. E. (2002). *Culture matters: How values shape human progress*. Basic books.

Heuser, R. (2016). Word vectors in the eighteenth century. *IPAM workshop: Cultural Analytics*.

Heuser, R. and Le-Khac, L. (2011). Learning to read data: Bringing out the humanistic in the digital humanities. *Victorian Studies*, 54(1):79-86.

Huxley, T. (1859). Darwin on the origins of species. *The Times*, 26 December: 8-9.

Jatowt, A. and Duh, K. (2014). A framework for analyzing semantic change of words across time. *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference*, 229-238. IEEE.

Jelveh, Z., Kogut, B., & Naidu, S. (2014). Detecting latent ideology in expert text: Evidence from academic papers in economics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1804-1809.

Kelly, B., P. D. S. A. and Taddy, M. (2017). Measuring technological innovation over the long run. *Working paper.*

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. *arXiv preprint* arXiv:1405.3515.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*, pages 625-635. International World Wide Web Conferences Steering Committee.

Landes, D. (2000). Culture makes almost all the difference. *Culture matters: how values shape human progress*, 2-13.

Lansley, C. M. (2016). Charles Darwin-s debt to the Romantics. *PhD thesis, University of Winchester*.

Levy, O., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Trans. ACL, 3.*

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (2):302-308.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the Google books ngram corpus. Pro*ceedings of the ACL 2012 system demonstrations*: 169-174. Association for Computational Linguistics.

Manovich, L. (2009). *Cultural analytics: visualising cultural patterns in the era of more media*. Domus March.

Marshall, A. (1890). Principles of political economy. Maxmillan, New York.

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

Mayr, E. (1995). Darwin's impact on modern thought. *Proceedings of the American Philosophical Society*, 139(4):317-325.

Mayr, E. (2001). The philosophical foundations of Darwinism. Proceedings of the American Philosophical Society, 145(4), 488-495.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176-182.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*: 3111-3119.

Mokyr, J. (2013). Cultural entrepreneurs and the origins of modern economic growth. *Scandinavian Economic History Review*, 61(1): 1-33.

Mokyr, J. (2016). A culture of growth: the origins of the modern economy. Princeton University Press.

Moretti, F. (2013). *Distant reading*. Verso Books.

Otis, L. (2009). Literature and science in the nineteenth century: an anthology. Oxford University Press.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10).

Richards, R. J. (2013). The impact of German romanticism on biology in the nineteenth century. The impact of Idealism: The legacy in philosophy and science, Cambridge University Press.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, *98*(5, Part 2), S71-S102.

Roth, S. (2014). Fashionable functions: A Google ngram view of trends in functional differentiation (1800-2000). *International Journal of Technology and Human Interaction*, 10(2):35-58.

Scholnick, R. (2015). American literature and science. University Press of Kentucky.

Sen, A. (2004). How does culture matter? In Rao, V. (2004). *Culture and public action*. Orient Blackswan.

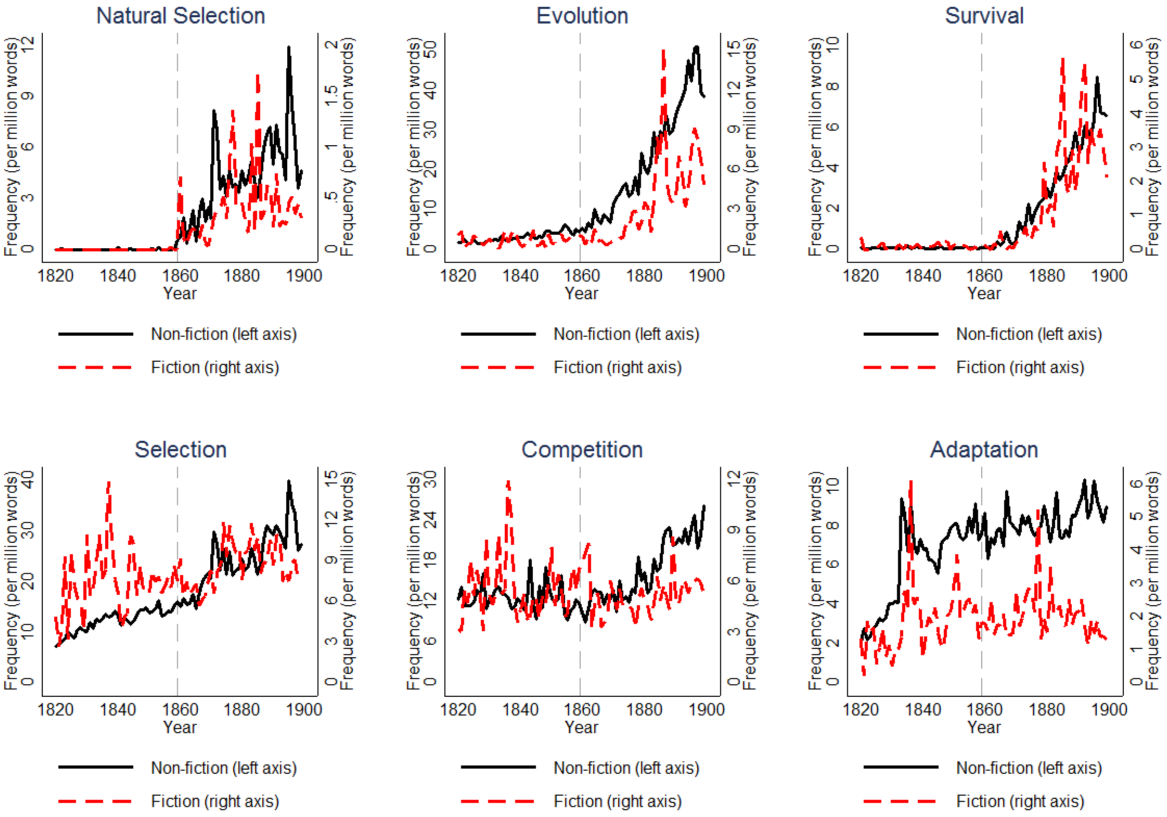
Stephan, P. E. (2012). *How economics shapes science* (Vol. 1). Cambridge, MA: Harvard University Press.

Turner, F. (2010). *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. University of Chicago Press.

Wilkens, M. (2015). Digital humanities and its application in the study of literature and culture. *Comparative Literature*, 67(1):11-20.

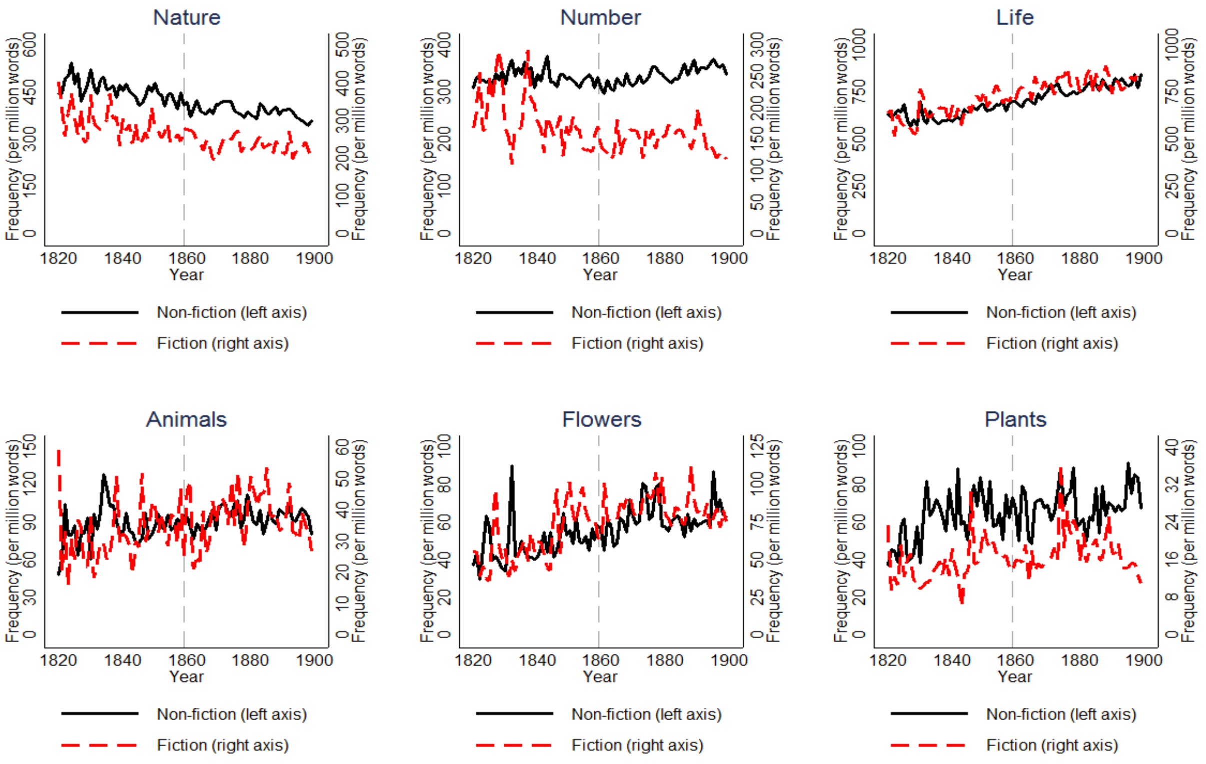
Winans, R. B. (1975). The Growth of a Novel-Reading Public in Late-Eighteenth-Century America. Early *American Literature*, 9(3), 267-275.

Figure 1: Frequencies (per 1 Million Words) of Selected Darwinian Words in the Google Books Corpora



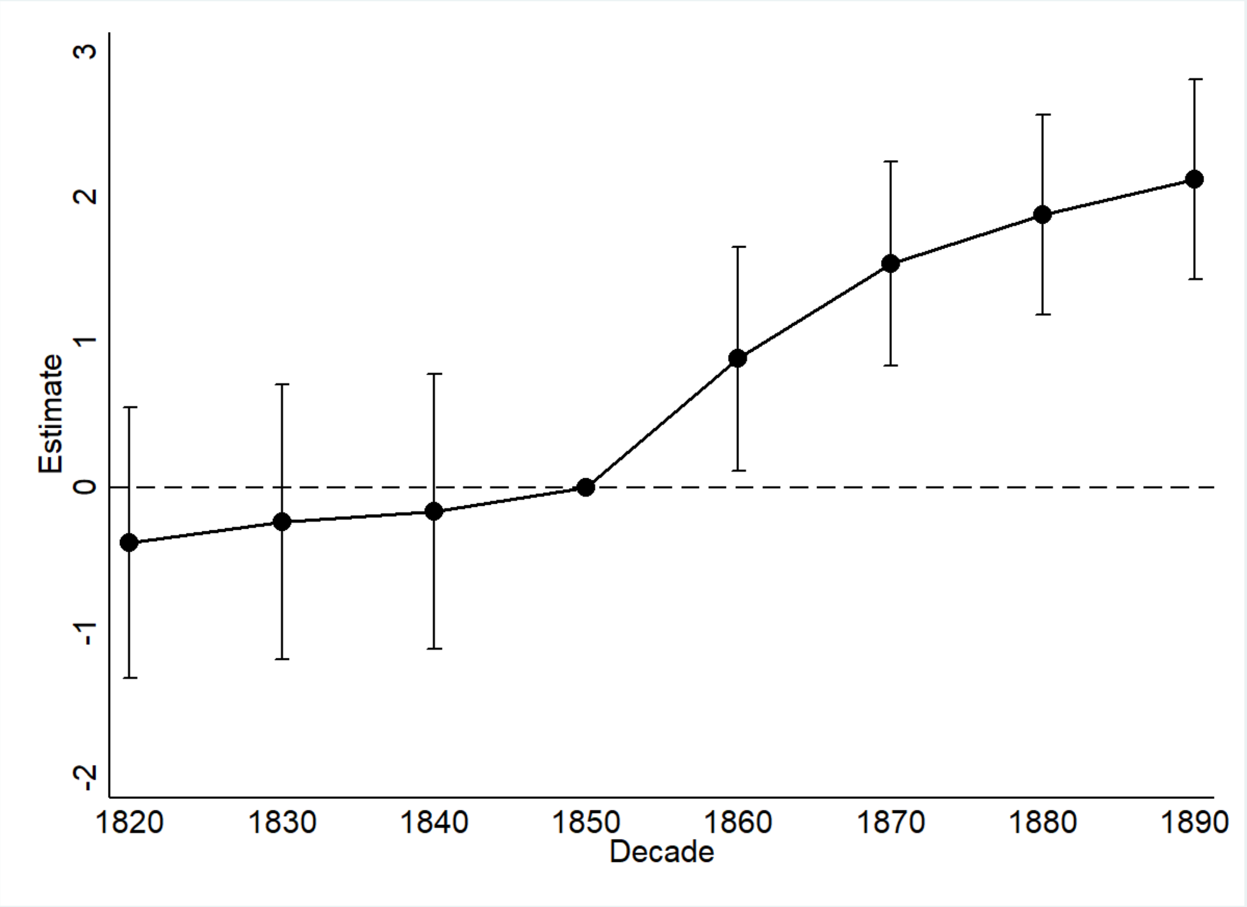
Notes: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words, separately for fiction and nonfiction texts. The y-axis on the left of each graph reports the reference scale for nonfiction, whereas the y-axis on the right shows the scale for fiction. Note that also the denominators for the calculation of the relative frequencies are separate for fiction and non-fiction.

Figure 2: Frequencies (per 1 Million Words) of Select "Generic" Words in the Google Books Corpora



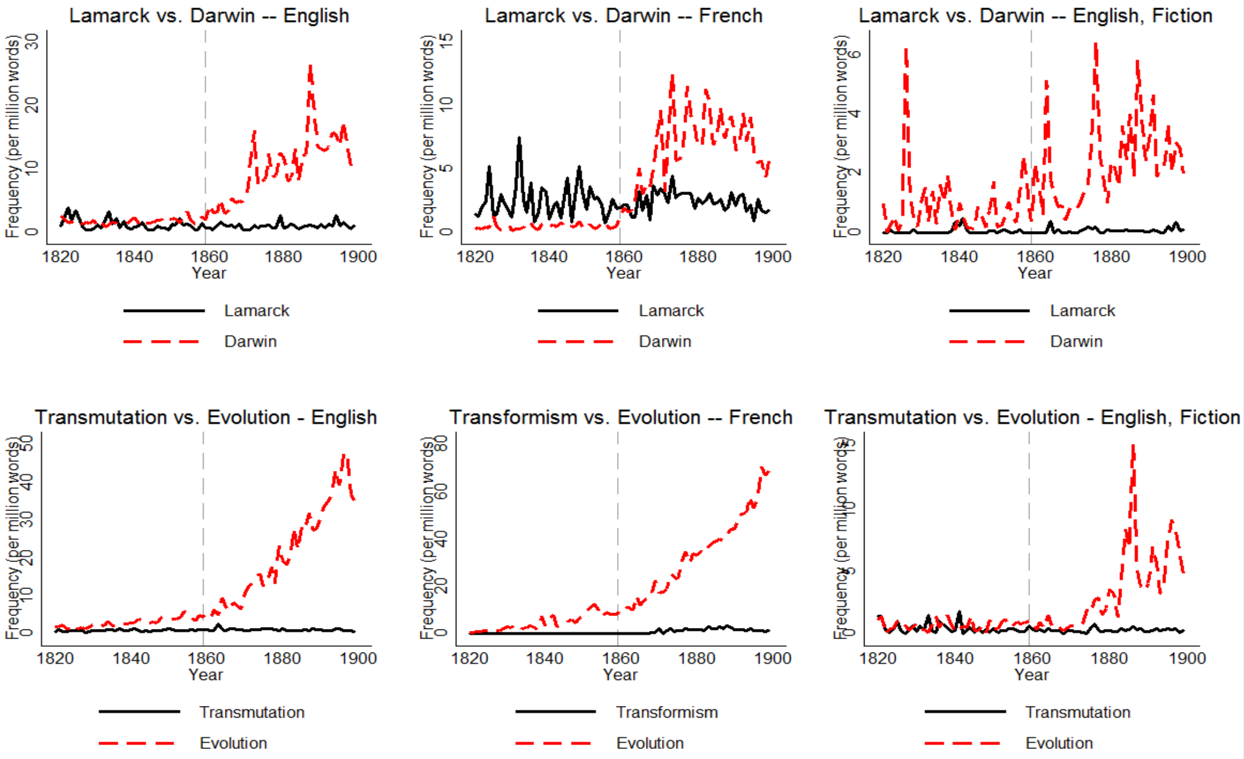
Notes: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words, separately for fiction and nonfiction texts. The y-axis on the left of each graph reports the reference scale for nonfiction, whereas the y-axis on the right shows the scale for fiction. Note that also the denominators for the calculation of the relative frequencies are separate for fiction and non-fiction.

Figure 3: Differences-in-Differences estimates of the average frequency of Darwinian and generic concepts in each decade between 1820 and 1899



Notes**:** Each dot in the graph represents the estimate of the parameters δ_j from the following regression model: $\ln\,(y_{wt}) = \alpha_w + \beta_w \mathbf{1}(Darwinian\ word) + \sum_{j=2}^{4} \gamma_i \mathbf{1}(\,j0 \leq t \leq j9) + \sum_{j=6}^{9} \gamma_j \mathbf{1}(j0 \leq t \leq j9) + \sum_{i=2}^{4} \delta_i \mathbf{1}(j0 \leq t \leq j9) * \mathbf{1}(Darwinian\ word) + \sum_{j=6}^{9} \delta_j \mathbf{1}(j0 \leq t \leq j9) * \mathbf{1}(Darwinian\ word) + \varepsilon_{wt}$, where y_{wt} is the frequency of use of a word per million words used (plus 0.01) and the omitted (or baseline) decade is 1850-59. The vertical bars report 95% confidence intervals (from robust standard errors). On the x-axis, 1820 represents the decade 1820-29, 1830 represents the decade 1830-39, and so on.

Figure 4: Frequencies (per 1 Million Words) of the Words “Lamarck” and “Darwin”, and “Transmutation” and “Evolution” in the English and French Google Books Corpora


Lamarck vs. Darwin -- English

Lamarck vs. Darwin -- French

Lamarck vs. Darwin -- English, Fiction

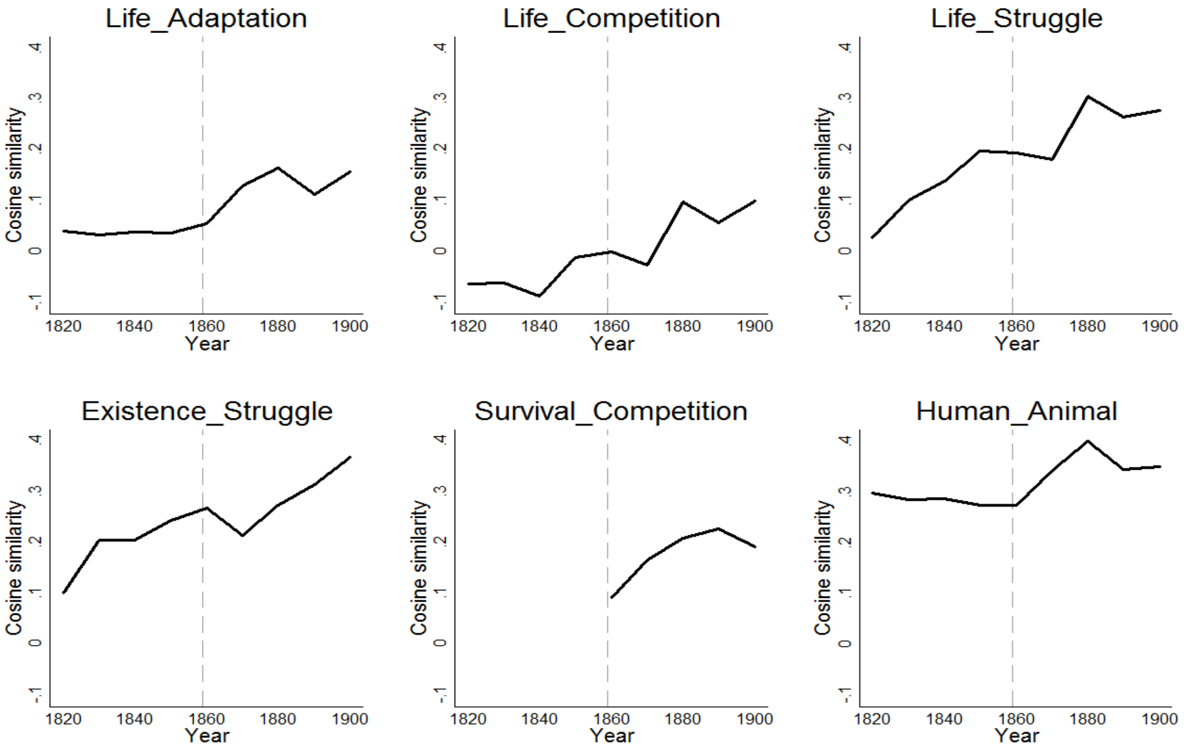
Transmutation vs. Evolution - English

Transformism vs. Evolution -- French

Transmutation vs. Evolution - English, Fiction

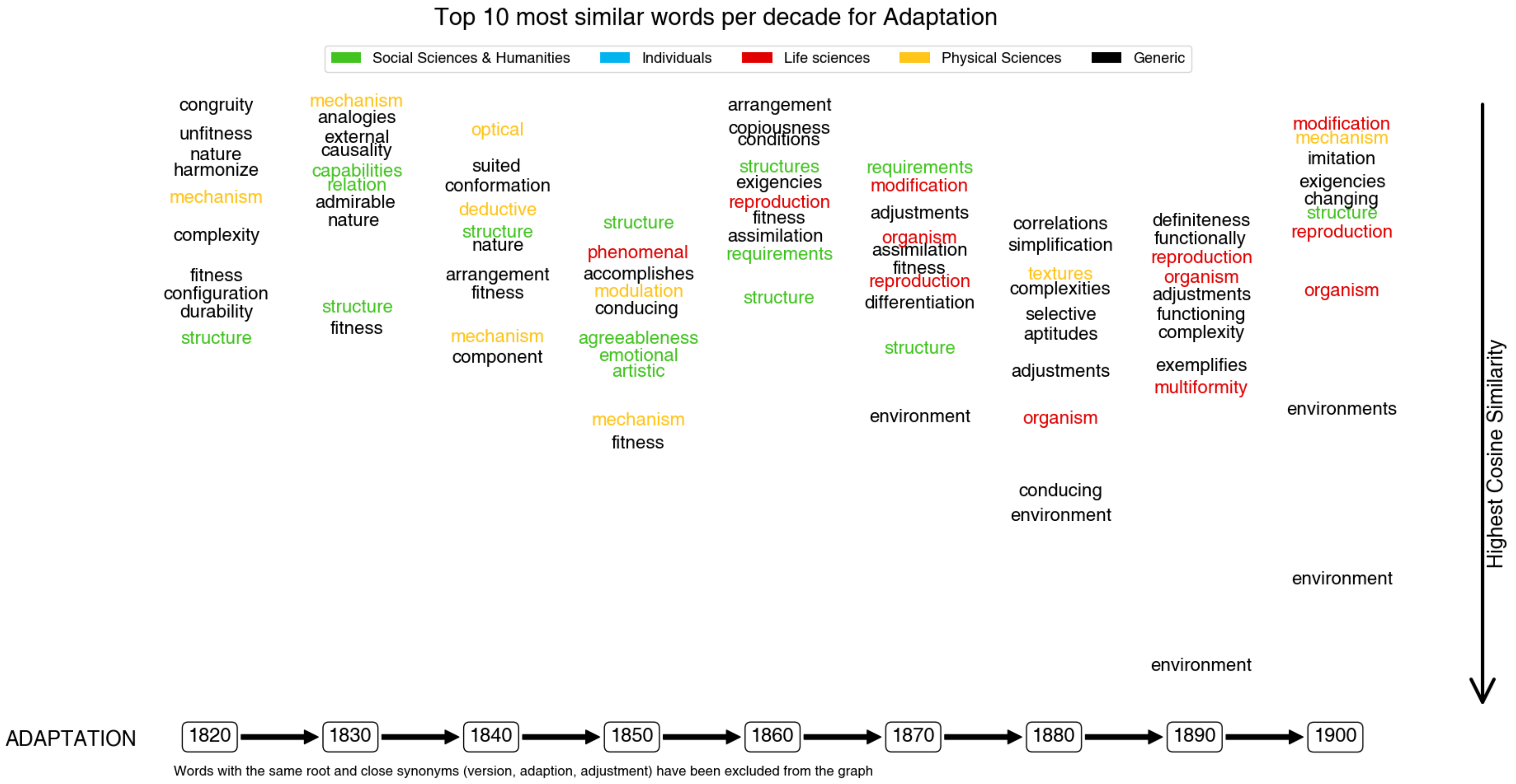
Notes: For each year, the figures report the number of occurrences (per million words) of the word or phrase indicated on top of a graph.

Figure 5: Semantic Associations between Selected Pairs of Words



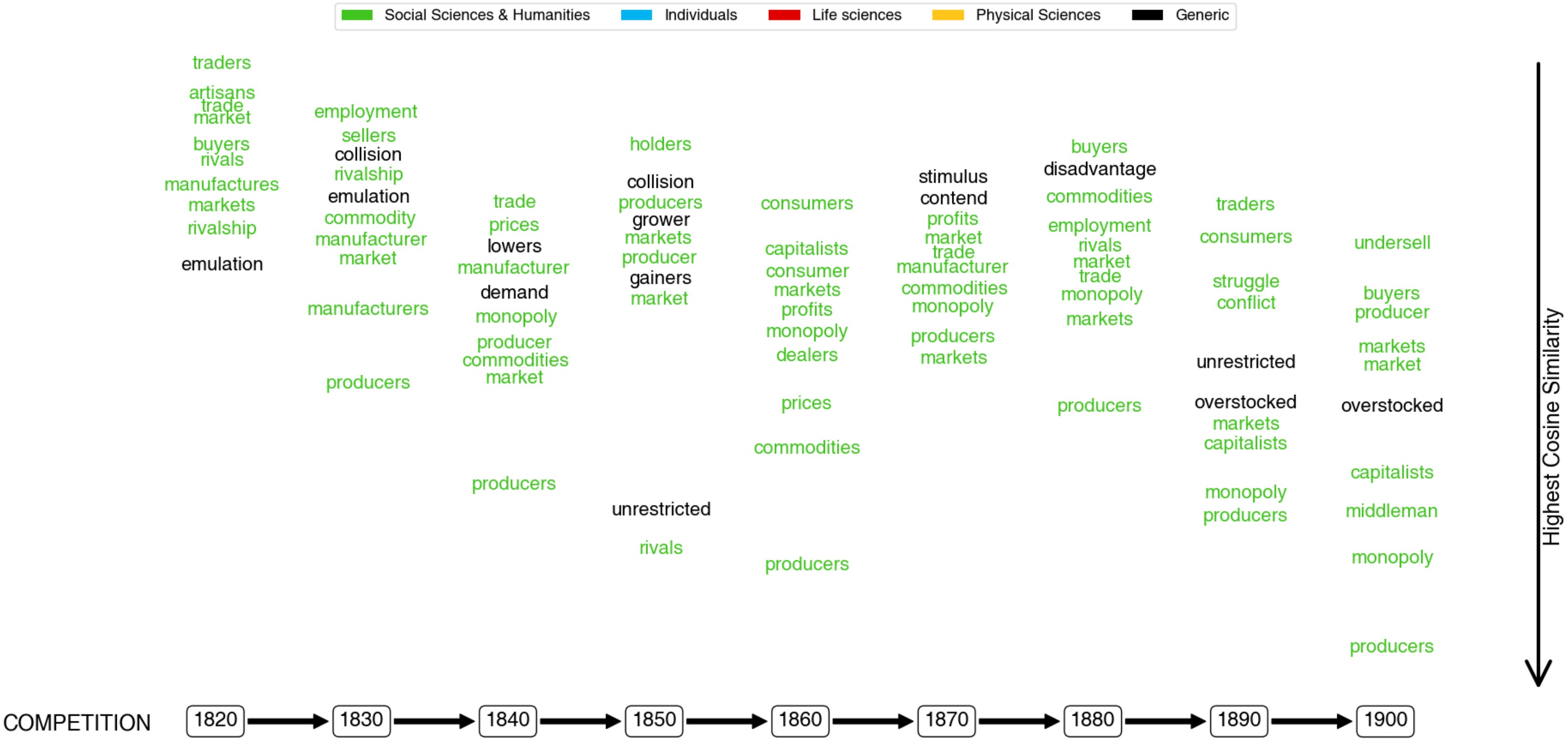
Notes: The graphs below report the similarity between each pair of words, as measured by the cosine of the angle between each pair of word vectors. The weights in the word vectors were calculated with a Word2Vec algorithm.

Figures 6 through 11: Top 10 most similar words for selected Darwinian words

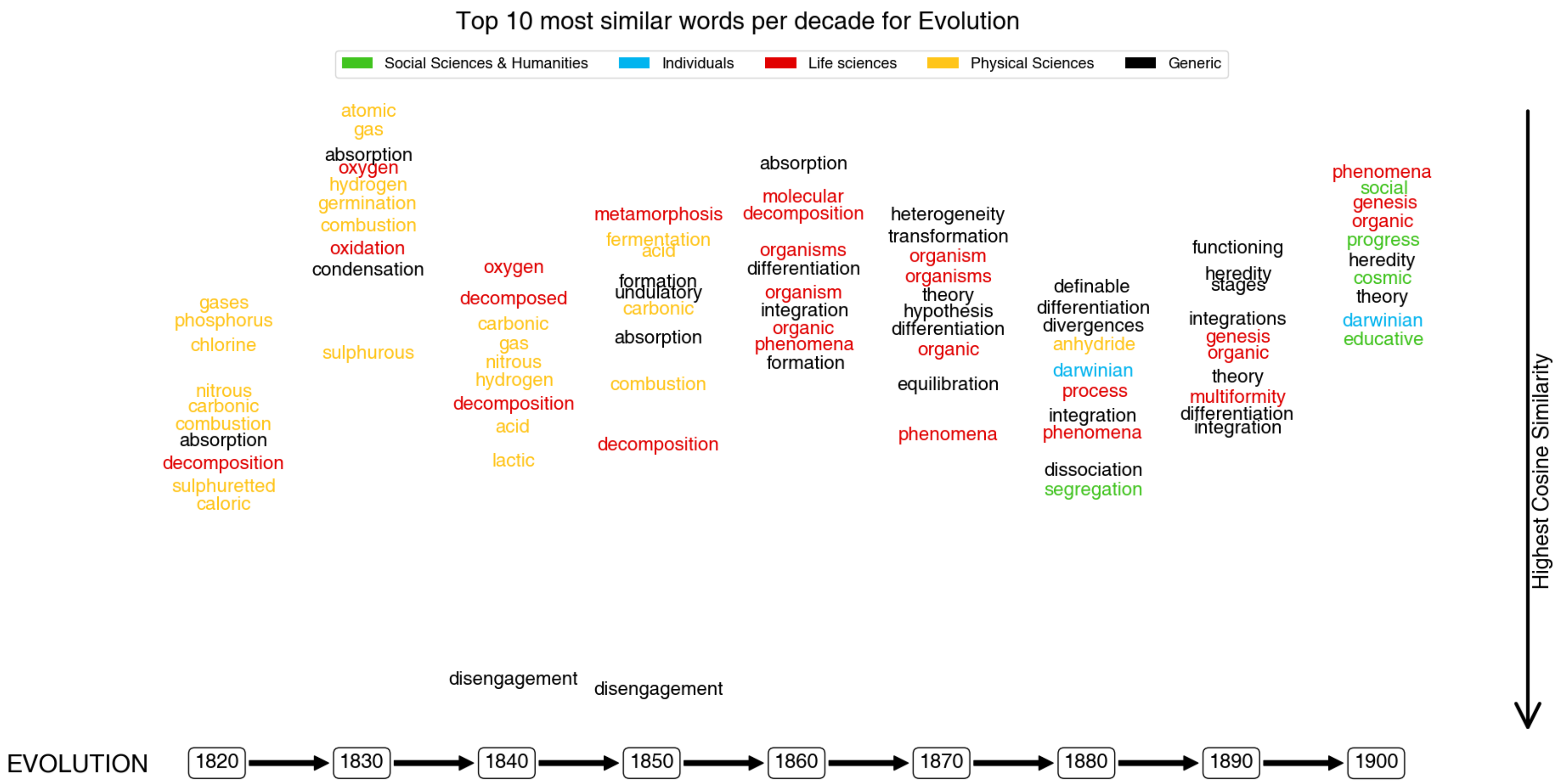


Words with the same root and close synonyms (version, adaption, adjustment) have been excluded from the graph

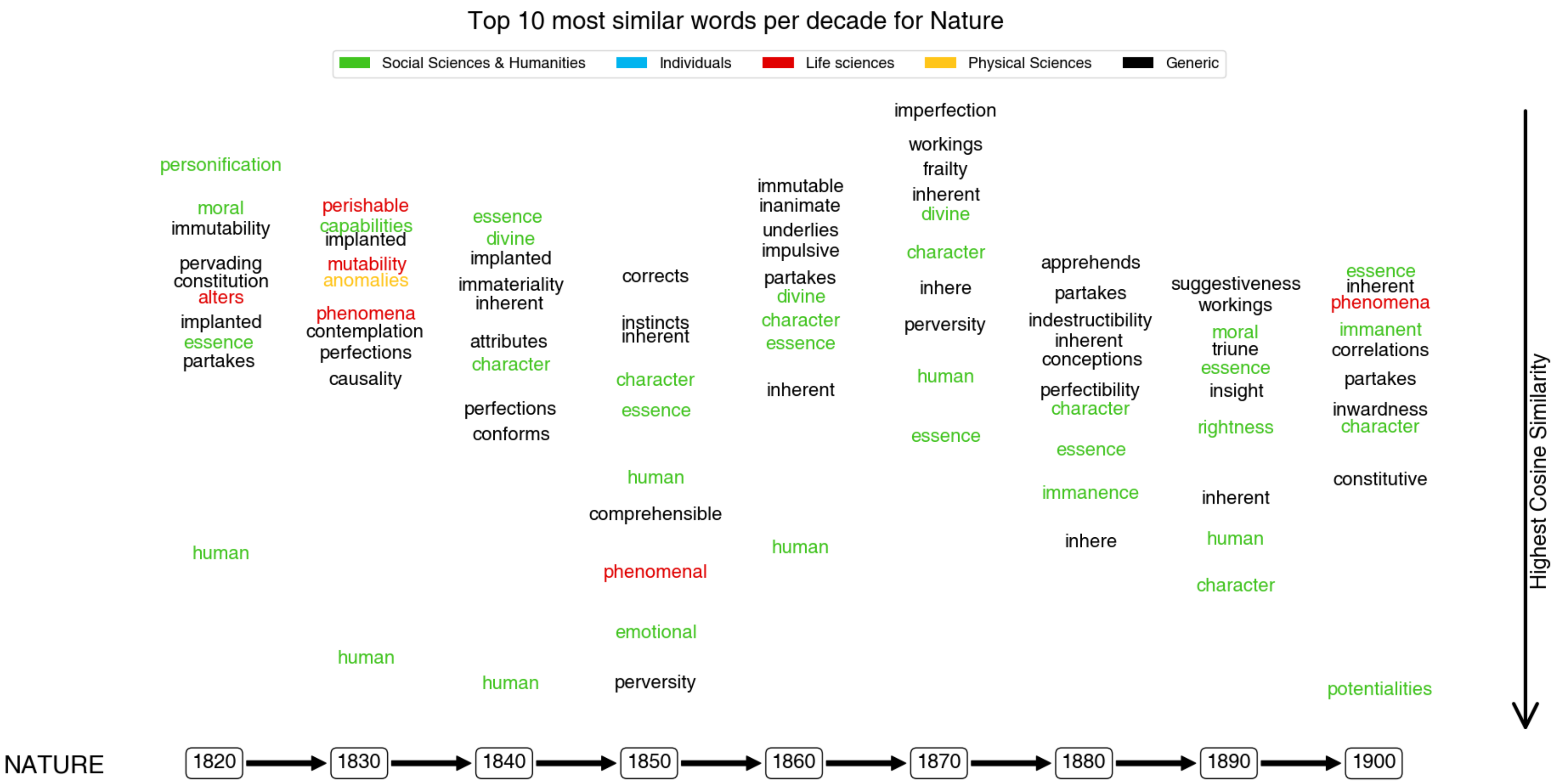
Top 10 most similar words per decade for Competition



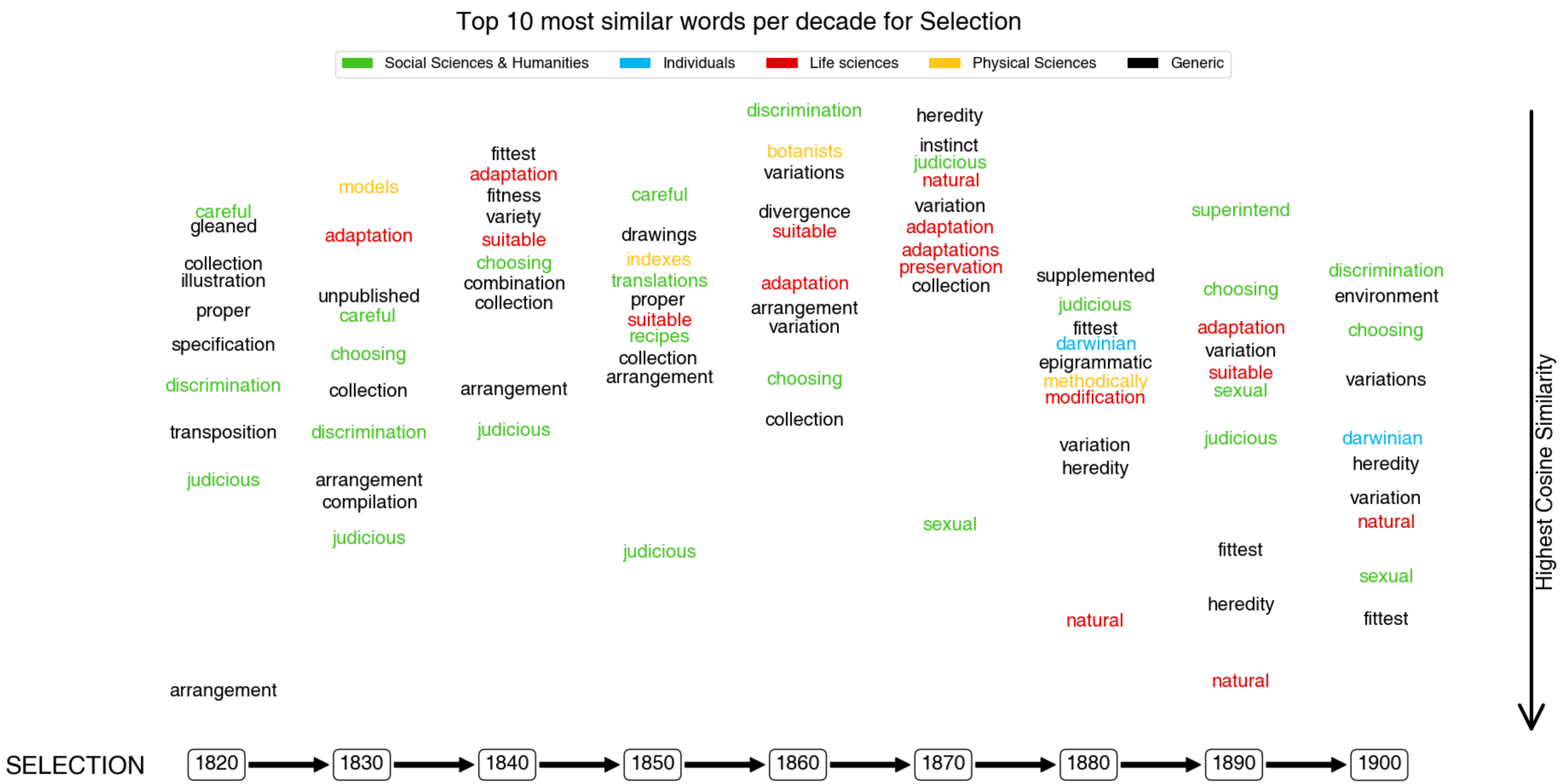
Words with the same root and close synonyms (contest, contention, rivalry, rival, challenger, competitor, contender) have been excluded from the graph



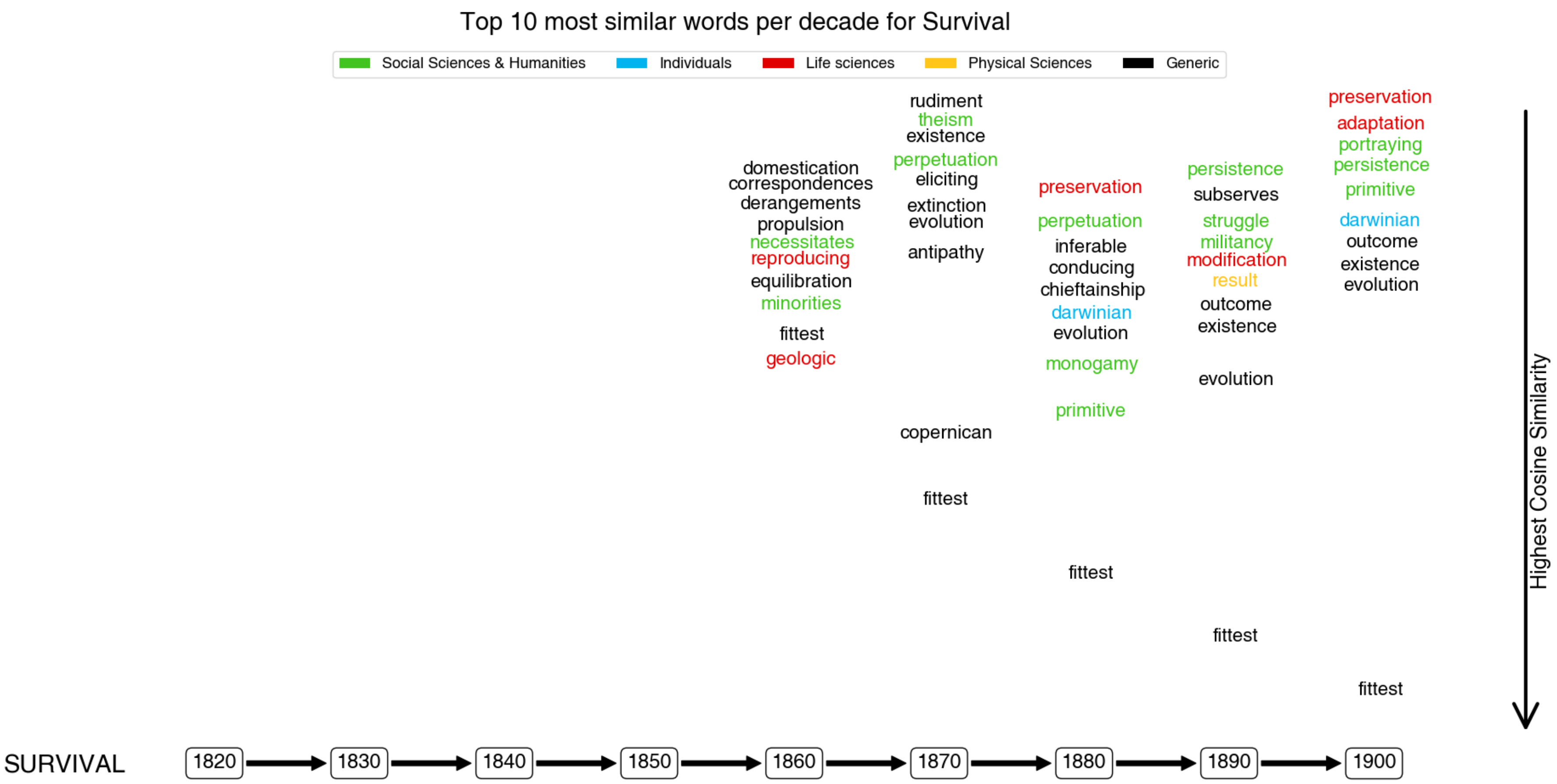
Words with the same root and close synonyms (development, phylogeny, phylogenesis) have been excluded from the graph



Words with the same root and close synonyms have been excluded from the graph


Top 10 most similar words per decade for Selection

Words with the same root and close synonyms (choice, option, pick, choice, pick, survival, excerpt, excerption, extract) have been excluded from the graph



Words with the same root and close synonyms (endurance, selection) have been excluded from the graph

Table 1: Regression Analyses – Frequency of Darwinian Concepts and Select Generic Words

A. Levels

Words:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection	Nature	Number	Fertility	Animals	Flowers	Plants
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year	0.087***	0.165***	-0.023	0.000	0.143***	0.002	-1.714***	-0.291	-0.045**	0.226	0.325***	0.578***
	(0.009)	(0.014)	(0.022)	(0.000)	(0.014)	(0.002)	(0.388)	(0.179)	(0.018)	(0.186)	(0.117)	(0.140)
1(Year>1859)	-4.888***	0.692	-2.616***	-1.120***	-1.408***	1.376**	-15.230*	-16.514***	-0.821**	0.600	4.039	-8.302*
	(1.102)	(0.999)	(0.793)	(0.190)	(0.358)	(0.542)	(9.046)	(5.185)	(0.385)	(3.885)	(3.273)	(4.609)
1(Year>1859) x (Year-1859)	0.958***	0.194***	0.343***	0.192***	-0.114***	0.137***	0.655	1.248***	0.010	-0.173	-0.124	-0.416**
	(0.060)	(0.050)	(0.032)	(0.009)	(0.017)	(0.028)	(0.453)	(0.234)	(0.021)	(0.215)	(0.157)	(0.192)
Constant	-0.420	5.549***	13.200***	0.056***	0.231	-0.068	518.948***	338.449***	11.297***	75.653***	36.668***	36.674***
	(0.329)	(0.596)	(0.821)	(0.017)	(0.673)	(0.057)	(17.365)	(8.011)	(0.828)	(8.956)	(5.662)	(5.820)
Observations	80	80	80	80	80	80	80	80	80	80	80	80
R-squared	0.961	0.889	0.752	0.969	0.694	0.795	0.774	0.274	0.624	0.094	0.445	0.247

B. Natural logarithms

Words:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection	Nature	Number	Fertility	Animals	Flowers	Plants
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year	1.094***	0.565***	-0.071	0.282	1.196***	0.860*	-0.135***	-0.023	-0.151**	0.162	0.271***	0.433***
	(0.121)	(0.043)	(0.066)	(0.209)	(0.086)	(0.460)	(0.034)	(0.021)	(0.070)	(0.099)	(0.088)	(0.087)
1(Year>1859)	0.015	0.031	-0.199***	0.404*	-0.254***	4.034***	-0.047**	-0.066***	-0.108**	-0.012	0.066	-0.140*
	(0.084)	(0.050)	(0.060)	(0.228)	(0.055)	(0.389)	(0.022)	(0.016)	(0.047)	(0.048)	(0.056)	(0.073)
1(Year>1859) x (Year-1859)	3.465***	0.715***	1.659***	8.471***	-0.902***	2.792***	-0.078	0.261***	-0.235*	-0.093	0.008	-0.228
	(0.230)	(0.148)	(0.132)	(0.603)	(0.134)	(0.798)	(0.061)	(0.041)	(0.140)	(0.138)	(0.155)	(0.185)
Constant	-2.944***	0.423***	2.757***	-3.599***	-2.651***	-7.332***	6.598***	5.871***	2.794***	3.837***	2.899***	2.490***
	(0.454)	(0.160)	(0.232)	(0.773)	(0.326)	(1.577)	(0.126)	(0.078)	(0.264)	(0.375)	(0.336)	(0.319)
Observations	80	80	80	80	80	80	80	80	80	80	80	80
R-squared	0.972	0.925	0.707	0.950	0.795	0.952	0.774	0.283	0.642	0.148	0.496	0.324

Notes: The tables report estimates from regressions of the annual frequency of use of a given word or phrase on a linear time trend, indicators for the years after 1859, and the interactions of these indicators with the difference between the current year and 1859. Each regression is limited to one word or phrase as indicated in the corresponding column, and includes 80 observations, one for each year from 1820 to 1899. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Frequencies are per million words; in absolute terms in Panel A, and as ln(frequency per million words + 0.01) in panel B. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Regression Analyses – Frequency of Darwinian Concepts: Fiction and Non-fiction

Word:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)
ln(Year)	1.118***	0.578***	-0.063	0.281	1.218***	0.861*
	(0.123)	(0.043)	(0.067)	(0.208)	(0.088)	(0.462)
1(Year>1859)	0.013	0.029	-0.206***	0.420*	-0.263***	4.036***
	(0.085)	(0.050)	(0.062)	(0.241)	(0.057)	(0.390)
1(Year>1859) x (ln(Year)-ln(59))	3.500***	0.776***	1.743***	8.560***	-0.850***	2.798***
	(0.230)	(0.149)	(0.139)	(0.629)	(0.140)	(0.801)
1(Fiction)	0.246	0.149	-1.483**	-2.700	-0.199	1.853
	(2.544)	(0.671)	(0.650)	(2.575)	(1.257)	(1.652)
ln(Year) x 1(Fiction)	-0.527	-0.191	0.173	0.632	-0.290	-0.601
	(0.672)	(0.176)	(0.174)	(0.695)	(0.332)	(0.482)
1(Year>1859) x 1(Fiction)	-0.627**	-0.161	0.048	-1.009	0.100	-1.400***
	(0.294)	(0.099)	(0.156)	(0.607)	(0.154)	(0.464)
1(Year>1859) x (ln(Year)-ln(59)) x 1(Fiction)	1.946**	-0.621**	-1.606***	0.270	-0.273	-0.834
	(0.842)	(0.281)	(0.403)	(1.473)	(0.440)	(1.021)
Constant	-2.995***	0.392**	2.752***	-3.609***	-2.699***	-7.335***
	(0.461)	(0.159)	(0.234)	(0.769)	(0.332)	(1.586)
Observations	160	160	160	160	160	160
R-squared	0.870	0.857	0.858	0.829	0.843	0.943

Notes: The table reports regressions of the natural logarithm of the relative annual frequency (per million words + 0.01) of use of a given word or phrase on the natural logarithm of a linear time trend, indicators for the years after 1859, the interactions of this indicators with the difference between the natural logarithm of the current year and the natural logarithm of 59, and interactions of all these previous terms with an indicator for whether an observation pertains to fiction books as opposed to non-fiction books. There are two observations per year, one based on the corpus of non-fiction books, and the other on the corpus of non-fiction books (N=160). Each regression is limited to one word or phrase as indicated in the corresponding columns. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

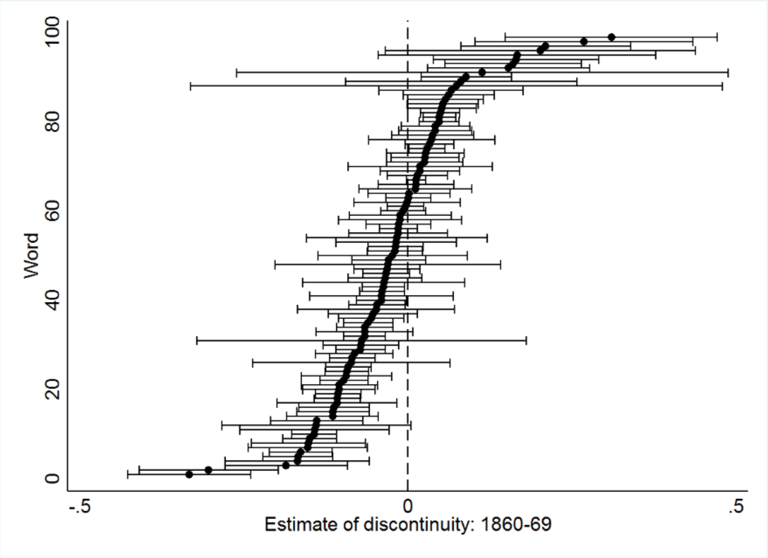
Table 3: Differences-in-Differences regressions – Darwinian and Generic Scientific Concepts

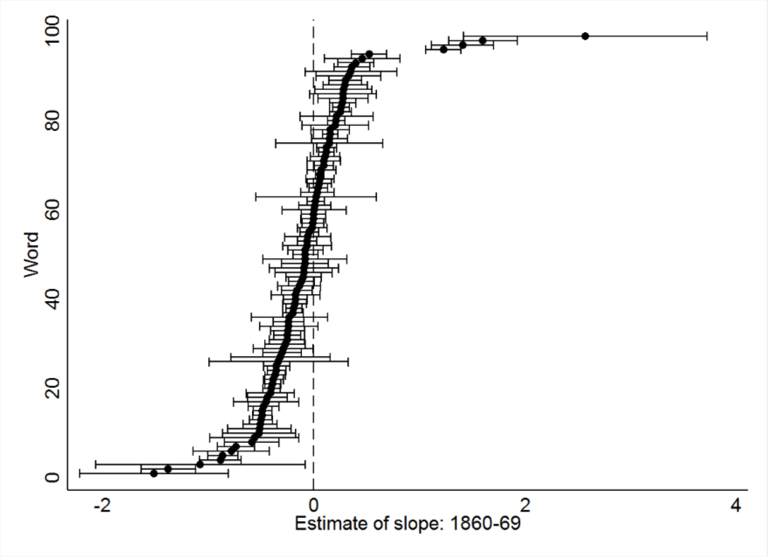
Outcome variable:	ln(aggregate frequency)	ln(aggregate frequency)	ln(aggregate frequency)
Sample:	Generic words	Darwinian words	Darwinian and generic words
	(1)	(2)	(3)
Regressors:			
ln(Year)	0.042***	0.405***	0.042***
	(0.010)	(0.035)	(0.010)
1(Year>1859) x ((ln(Year)-ln(59))	-0.021	1.723***	-0.021
	(0.020)	(0.092)	(0.020)
1(Darwinian word			-7.471***
			(0.134)
1(Darwinian word) x ln(Year)			0.363***
			(0.036)
1(Darwinian word) x 1(Year>1859) x ((ln(Year)-ln(59))			1.744***
			(0.095)
Constant	9.482***	2.011***	9.482***
	(0.038)	(0.129)	(0.038)
Observations	80	80	160
R-squared	0.407	0.966	1.000

Notes: Columns 1 and 2 report estimates from regressions where the outcome variable is the natural logarithm of the aggregate frequency of the 99 most frequent nouns in *On the Origins of* Species (column 1) and of the aggregate yearly frequencies of the six Darwinian word and concepts (column 2). The regression estimates in column 3 come from combining the data used for the regressions in columns 1 and 2; therefore there are two observations per year (N=160). Robust standard errors are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Appendix

Figure A1: Estimates of slopes and discontinuities in the frequency of 99 high-frequency nouns in the 1860-69 decade





Notes: Each dot represents the estimates of parameters $\gamma_w$and δ_w (top and bottom graph respectively) from regression model 4, for the 99 high frequency generic nouns. Each word is represented by a number between 1 and 99 on the vertical axis (Table A1 reports the list of these words). The horizontal lines and bars are the 95% confidence intervals of the estimates.

Table A1: Generic Words

action
advantage
animal
animals
beings
birds
breeds
case
cases
change
changes
character
characters
class
climate
conditions
country
degree
descendants
descent
development
difference
differences
difficulty
eggs
fact
facts
fertility
flower
flowers
form
formation
formations
forms
genera
generations
genus
group
groups
habits
hand
hybrids
importance
individuals
inhabitants
insects
instance
instincts
islands
kind
kinds
land
life
man
manner
means
modification
naturalists
nature
number
numbers
offspring
older
organ
organization
organs
parent
parts
period
periods
plant
plants
points
pollen
power
principle
process
productions
reason
respect
sea
seeds
size
species
state
sterility
structure
subject
tendency
theory
time
variation
variations
varieties
variety
view
water
world
years

Notes: The table lists the 99 most frequent nouns in *On the Origins of Species*, which we used as controls for the Darwinian concepts in some of the analyses.

Table A2: Regression Analyses – Frequency of Darwinian Concepts and Select Generic Words: Full set of indicators

A. Levels

Word:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection	Nature	Number	Fertility	Animals	Flowers	Plants
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Year	-0.029	0.369***	0.097	0.003	0.169***	-0.001*	-1.272	2.076*	-0.194	2.275	-0.006	0.875
	(0.034)	(0.055)	(0.212)	(0.003)	(0.034)	(0.000)	(3.299)	(1.099)	(0.169)	(1.618)	(0.796)	(0.692)
1(Year>1829)	0.754**	-0.414	0.705	-0.012	1.030	0.005**	26.915	7.730	1.890	0.070	11.494	6.780
	(0.286)	(0.595)	(1.415)	(0.023)	(1.175)	(0.002)	(28.114)	(15.591)	(1.638)	(15.384)	(14.766)	(11.794)
1(Year>1839)	0.175	-2.674***	-0.443	-0.048**	-2.258***	0.030	2.148	-0.541	-0.667	-17.569	0.531	-10.022
	(0.316)	(0.748)	(1.489)	(0.023)	(0.845)	(0.022)	(16.027)	(18.461)	(0.933)	(11.769)	(4.813)	(9.965)
1(Year>1849)	0.280	-0.598	-1.436	0.025	0.601	-0.063	23.329	-2.249	0.737	8.150	-2.517	-6.668
	(0.531)	(0.663)	(1.584)	(0.021)	(0.397)	(0.085)	(19.220)	(8.097)	(0.762)	(5.748)	(5.118)	(7.984)
1(Year>1859)	0.478	-1.244	-0.435	-0.042	-0.622	0.526	-8.155	-19.921**	-0.853	-5.027	-3.902	-2.513
	(0.895)	(0.962)	(1.053)	(0.078)	(0.775)	(0.477)	(18.287)	(8.371)	(0.876)	(5.972)	(5.472)	(8.884)
1(Year>1869)	2.427	4.979*	-1.819	0.054	-0.465	2.987	36.551***	-12.017	-0.554	11.375*	0.868	0.061
	(1.513)	(2.883)	(1.443)	(0.271)	(0.572)	(2.072)	(13.735)	(7.838)	(0.568)	(6.094)	(7.201)	(7.109)
1(Year>1879)	0.150	-1.283	-2.429	-0.220	-0.443	-0.238	0.730	-33.799***	-0.487	-6.592	-13.832**	-19.936**
	(2.973)	(1.678)	(1.838)	(0.269)	(0.551)	(0.900)	(12.129)	(7.499)	(0.515)	(8.249)	(6.569)	(9.032)
1(Year>1889)	3.087	1.341	-2.418*	-0.202	1.261**	0.351	7.423	-5.912	-0.398	5.405	-1.760	-5.601
	(3.163)	(2.059)	(1.288)	(0.389)	(0.494)	(1.232)	(8.615)	(8.140)	(0.503)	(4.943)	(3.477)	(5.816)
1(Year>1829) x (Year-1829)	0.048	-0.012	-0.287	0.001	0.176	0.000	-3.479	-2.486	0.189	-0.929	-1.287	0.326
	(0.040)	(0.083)	(0.232)	(0.004)	(0.161)	(0.000)	(4.454)	(2.820)	(0.272)	(2.727)	(1.905)	(1.705)
1(Year>1839) x (Year-1839)	0.099	-0.072	0.457**	-0.004	-0.222	-0.003	3.102	-1.199	-0.124	-1.787	2.740	0.120
	(0.064)	(0.115)	(0.227)	(0.003)	(0.165)	(0.003)	(3.665)	(2.934)	(0.217)	(2.350)	(1.823)	(1.891)
1(Year>1849) x (Year-1849)	-0.076	-0.161	-0.433*	-0.002	-0.136*	0.029	-3.295	1.491	0.017	0.605	-1.521*	-1.557
	(0.082)	(0.131)	(0.233)	(0.004)	(0.076)	(0.022)	(3.096)	(1.715)	(0.131)	(0.971)	(0.792)	(1.430)
1(Year>1859) x (Year-1859)	0.230*	0.515***	0.417**	0.045**	0.115	0.143**	3.442	2.096	0.184	0.215	1.525	0.726
	(0.130)	(0.154)	(0.159)	(0.021)	(0.119)	(0.071)	(3.072)	(1.469)	(0.142)	(1.032)	(0.964)	(1.280)
1(Year>1869) x (Year-1869)	0.528	-0.940**	0.144	0.151***	-0.045	-0.354	-3.043	1.747	-0.123	-0.504	-0.697	0.390
	(0.386)	(0.390)	(0.254)	(0.034)	(0.117)	(0.275)	(2.517)	(1.344)	(0.105)	(1.354)	(1.315)	(1.587)
1(Year>1879) x (Year-1879)	0.420	1.016**	0.487	0.087*	-0.043	0.484*	5.638***	-0.037	0.110	0.070	-0.434	0.517
	(0.466)	(0.424)	(0.315)	(0.049)	(0.077)	(0.284)	(2.009)	(1.287)	(0.098)	(1.239)	(1.100)	(1.474)
1(Year>1889) x (Year-1889)	-0.351	-0.766*	-0.577**	-0.077	-0.100	-0.359*	-6.673***	-4.275***	-0.074	-0.914	0.686	-0.153
	(0.686)	(0.388)	(0.279)	(0.079)	(0.079)	(0.213)	(1.625)	(1.469)	(0.094)	(1.025)	(0.577)	(1.075)
Constant	2.441***	0.026	10.049**	-0.017	-1.286	0.015*	503.102***	273.148***	14.377***	20.763	44.161**	24.277
	(0.851)	(1.274)	(4.978)	(0.087)	(0.832)	(0.009)	(82.227)	(27.148)	(4.295)	(42.889)	(20.418)	(16.130)
Observations	80	80	80	80	80	80	80	80	80	80	80	80
R-squared	0.971	0.923	0.808	0.982	0.829	0.835	0.820	0.501	0.698	0.372	0.557	0.420

Notes: The table reports estimates from regressions of the relative annual frequency of use (per million words) of a given word or phrase on a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, and interactions of these indicators with the difference between the current year and 1829, 39, 49, 59, 69, 79 and 89, respectively. Each regression is limited to one word or phrase as indicated in the corresponding column, and includes 80 observations, one for each year from 1820 to 1899. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

B. Natural logarithms

Word:	Evolution	Selection	Competition	Survival	Adaptation	Natural_Selection	Nature	Number	Fertility	Animals	Flowers	Plants
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
ln(Year)	-0.409	1.020***	0.150	1.164	1.412***	-1.003*	-0.049	0.152*	-0.479	0.892	0.079	0.462
	(0.454)	(0.129)	(0.383)	(1.312)	(0.272)	(0.530)	(0.172)	(0.084)	(0.409)	(0.537)	(0.439)	(0.359)
1(Year>1829)	0.387***	-0.048	0.074	-0.148	0.153	0.367**	0.058	0.023	0.166	-0.065	0.169	0.088
	(0.139)	(0.056)	(0.105)	(0.298)	(0.188)	(0.165)	(0.060)	(0.049)	(0.175)	(0.192)	(0.276)	(0.230)
1(Year>1839)	0.090	-0.215***	-0.056	-0.537**	-0.357***	0.954	0.005	-0.003	-0.064	-0.202	-0.007	-0.179
	(0.101)	(0.065)	(0.110)	(0.265)	(0.127)	(0.694)	(0.035)	(0.055)	(0.093)	(0.133)	(0.095)	(0.156)
1(Year>1849)	0.078	-0.040	-0.095	0.342	0.091	0.319	0.057	-0.007	0.084	0.102	-0.039	-0.089
	(0.133)	(0.048)	(0.122)	(0.261)	(0.059)	(0.791)	(0.044)	(0.026)	(0.087)	(0.067)	(0.092)	(0.122)
1(Year>1859)	0.075	-0.064	-0.052	-0.021	-0.088	2.653**	-0.022	-0.065**	-0.113	-0.059	-0.067	-0.052
	(0.156)	(0.061)	(0.097)	(0.387)	(0.107)	(1.101)	(0.045)	(0.027)	(0.108)	(0.072)	(0.100)	(0.148)
1(Year>1869)	0.312**	0.219*	-0.152	0.271	-0.055	0.628	0.095***	-0.039	-0.073	0.130*	0.014	0.005
	(0.128)	(0.128)	(0.119)	(0.474)	(0.072)	(0.552)	(0.035)	(0.024)	(0.079)	(0.066)	(0.107)	(0.110)
1(Year>1879)	0.032	-0.051	-0.144	-0.116	-0.058	-0.071	-0.001	-0.103***	-0.066	-0.065	-0.203**	-0.304**
	(0.154)	(0.072)	(0.122)	(0.133)	(0.069)	(0.211)	(0.032)	(0.023)	(0.073)	(0.087)	(0.093)	(0.133)
1(Year>1889)	0.073	0.053	-0.124*	-0.060	0.153**	0.072	0.023	-0.016	-0.053	0.065	-0.031	-0.080
	(0.095)	(0.076)	(0.067)	(0.070)	(0.059)	(0.208)	(0.023)	(0.024)	(0.075)	(0.056)	(0.052)	(0.088)
1(Year>1829) x (ln(Year)-ln(29))	0.698	0.010	-0.671	0.579	0.924	0.071	-0.295	-0.184	0.571	-0.194	-0.669	0.416
	(0.552)	(0.229)	(0.464)	(1.572)	(0.857)	(0.839)	(0.274)	(0.282)	(0.862)	(1.044)	(1.161)	(1.039)
1(Year>1839) x (ln(Year)-ln(39))	1.106	-0.066	1.433**	-1.387	-1.564*	-3.812	0.164	-0.177	-0.722	-0.949	1.809	-0.009
	(0.820)	(0.411)	(0.717)	(1.600)	(0.888)	(4.027)	(0.301)	(0.323)	(0.788)	(0.998)	(1.169)	(1.198)
1(Year>1849) x (ln(Year)-ln(49))	-0.841	-0.511	-1.671**	-2.036	-0.884*	11.975	-0.449	0.185	-0.086	0.348	-1.298*	-1.093
	(1.011)	(0.489)	(0.825)	(2.583)	(0.528)	(9.622)	(0.352)	(0.252)	(0.792)	(0.534)	(0.709)	(1.066)
1(Year>1859) x (ln(Year)-ln(59))	2.452*	1.857***	2.280***	13.781***	0.986	0.701	0.381	0.437	1.344	0.173	1.613	0.810
	(1.353)	(0.574)	(0.837)	(4.332)	(0.996)	(9.453)	(0.442)	(0.276)	(0.978)	(0.741)	(1.011)	(1.235)
1(Year>1869) x (ln(Year)-ln(69))	0.734	-3.159**	0.616	-1.259	-0.358	-8.989	-0.613	0.445	-1.146	-0.433	-0.726	0.285
	(1.932)	(1.267)	(1.414)	(4.647)	(1.036)	(5.974)	(0.437)	(0.284)	(1.009)	(1.022)	(1.391)	(1.701)
1(Year>1879) x (ln(Year)-ln(79))	0.627	3.185**	2.197	-4.559	-0.354	5.575	1.118***	0.077	1.160	0.109	-0.380	1.016
	(1.813)	(1.379)	(1.675)	(2.899)	(0.779)	(5.103)	(0.413)	(0.310)	(1.073)	(1.018)	(1.193)	(1.671)
1(Year>1889) x (ln(Year)-ln(89))	-2.162	-2.598**	-3.001**	-2.922**	-1.118	-6.088*	-1.714***	-1.099***	-0.948	-1.016	1.030	-0.228
	(1.775)	(1.280)	(1.360)	(1.348)	(0.858)	(3.331)	(0.372)	(0.400)	(1.221)	(1.048)	(0.777)	(1.455)
Constant	1.839	-1.058**	2.034*	-6.472	-3.473***	-1.298	6.312***	5.296***	3.785***	1.470	3.513**	2.333**
	(1.448)	(0.404)	(1.205)	(4.226)	(0.880)	(1.755)	(0.547)	(0.269)	(1.312)	(1.756)	(1.416)	(1.126)
Observations	80	80	80	80	80	80	80	80	80	80	80	80
R-squared	0.979	0.955	0.765	0.970	0.903	0.963	0.836	0.507	0.710	0.407	0.582	0.462

Notes: The table reports regressions of the natural logarithm relative annual frequency of use (per million words +0.01) of a given word or phrase on the natural logarithm of a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, and interactions of these indicators with the difference between the natural logarithm of the current year and the natural logarithm of (18)29, 39, 49, 59, 69, 79 and 89, respectively. Each regression is limited to one word or phrase as indicated in the corresponding column, and includes 80 observations, one for each year from 1820 to 1899. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A3: Regression Analyses – Frequency of Darwinian Concepts: Fiction and Non-fiction: Full set of indicators

Word:	Evolution	Selection	Competition	Survival	Adaptation	Natural_Selection
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)
ln(Year)	-0.395	1.022***	0.132	1.163	1.448***	-1.007*
	(0.470)	(0.123)	(0.400)	(1.240)	(0.284)	(0.534)
1(Year>1829)	0.386***	-0.055	0.071	-0.139	0.149	0.369**
	(0.138)	(0.059)	(0.111)	(0.339)	(0.188)	(0.166)
1(Year>1839)	0.104	-0.200***	-0.046	-0.509*	-0.347***	0.958
	(0.102)	(0.067)	(0.112)	(0.285)	(0.128)	(0.700)
1(Year>1849)	0.081	-0.037	-0.100	0.417	0.084	0.321
	(0.135)	(0.051)	(0.124)	(0.288)	(0.060)	(0.795)
1(Year>1859)	0.072	-0.069	-0.065	-0.089	-0.095	2.654**
	(0.160)	(0.058)	(0.103)	(0.402)	(0.115)	(1.106)
1(Year>1869)	0.316**	0.222*	-0.158	0.281	-0.059	0.627
	(0.131)	(0.131)	(0.123)	(0.485)	(0.076)	(0.554)
1(Year>1879)	0.042	-0.040	-0.134	-0.114	-0.047	-0.071
	(0.155)	(0.076)	(0.126)	(0.140)	(0.075)	(0.211)
1(Year>1889)	0.094	0.066	-0.109	-0.073	0.167***	0.074
	(0.094)	(0.077)	(0.072)	(0.070)	(0.061)	(0.208)
1(Year>1829) x (ln(Year)-ln(29))	0.702	0.028	-0.632	0.575	0.916	0.073
	(0.553)	(0.245)	(0.492)	(1.571)	(0.857)	(0.844)
1(Year>1839) x (ln(Year)-ln(39))	1.056	-0.124	1.402*	-1.898	-1.616*	-3.833
	(0.824)	(0.428)	(0.737)	(1.826)	(0.899)	(4.060)
1(Year>1849) x (ln(Year)-ln(49))	-0.801	-0.453	-1.656**	-1.281	-0.782	12.002
	(1.024)	(0.489)	(0.831)	(2.730)	(0.556)	(9.679)
1(Year>1859) x (ln(Year)-ln(59))	2.587*	1.996***	2.500***	14.084***	1.023	0.725
	(1.407)	(0.551)	(0.894)	(4.430)	(1.058)	(9.499)
1(Year>1869) x (ln(Year)-ln(69))	0.395	-3.494***	0.334	-2.008	-0.599	-9.027
	(1.975)	(1.289)	(1.471)	(4.760)	(1.104)	(5.990)
1(Year>1879) x (ln(Year)-ln(79))	0.847	3.446**	2.303	-4.343	-0.153	5.591
	(1.806)	(1.438)	(1.736)	(2.994)	(0.847)	(5.122)
1(Year>1889) x (ln(Year)-ln(89))	-2.099	-2.535*	-2.955**	-2.466*	-1.014	-6.103*
	(1.741)	(1.319)	(1.451)	(1.372)	(0.921)	(3.336)

(continues to next page)

(continues from previous page)

Word:	Evolution	Selection	Competition	Survival	Adaptation	Natural_Selection
Regressors:	(1)	(2)	(3)	(4)	(5)	(6)
1(Fiction)	-5.557	-0.445	-5.165	-1.298	0.910	-3.321*
	(13.484)	(3.134)	(4.344)	(14.611)	(7.347)	(1.767)
ln(Year) x 1(Fiction)	1.312	-0.033	1.326	0.193	-0.607	1.007*
	(4.083)	(0.973)	(1.383)	(4.573)	(2.243)	(0.534)
1(Year>1829) x 1(Fiction)	-1.007	0.278	-0.188	-0.956	-0.780**	-0.369**
	(0.677)	(0.319)	(0.392)	(1.309)	(0.340)	(0.166)
1(Year>1839) x 1(Fiction)	-1.198**	-0.459*	-0.418	-1.384	-0.832**	-0.958
	(0.532)	(0.256)	(0.371)	(1.185)	(0.406)	(0.700)
1(Year>1849) x 1(Fiction)	-0.148	-0.119	0.217	-0.436	0.477**	-0.141
	(0.614)	(0.153)	(0.252)	(1.222)	(0.234)	(0.801)
1(Year>1859) x 1(Fiction)	-0.190	0.056	0.322	1.243	0.232	0.468
	(0.273)	(0.136)	(0.293)	(1.069)	(0.181)	(1.423)
1(Year>1869) x 1(Fiction)	-0.076	-0.062	0.383	0.700	0.522**	0.608
	(0.318)	(0.219)	(0.242)	(0.958)	(0.235)	(0.555)
1(Year>1879) x 1(Fiction)	0.045	-0.141	0.049	0.192	0.001	-0.792
	(0.457)	(0.169)	(0.192)	(0.551)	(0.424)	(0.499)
1(Year>1889) x 1(Fiction)	-0.423	-0.050	-0.069	0.442	-0.066	-0.701
	(0.389)	(0.136)	(0.240)	(0.392)	(0.193)	(0.429)
1(Year>1829) x (ln(Year)-ln(29)) x 1(Fiction)	2.832	-0.222	-0.395	6.384	4.363*	-0.073
	(4.768)	(1.609)	(2.129)	(5.775)	(2.568)	(0.844)
1(Year>1839) x (ln(Year)-ln(39)) x 1(Fiction)	-3.764	1.477	-0.885	1.010	-2.509	3.833
	(4.667)	(1.652)	(2.029)	(9.772)	(2.414)	(4.060)
1(Year>1849) x (ln(Year)-ln(49)) x 1(Fiction)	3.241	-1.275	0.635	-13.954	-4.732**	-11.146
	(4.212)	(1.210)	(1.923)	(10.286)	(2.348)	(11.008)
1(Year>1859) x (ln(Year)-ln(59)) x 1(Fiction)	-11.519***	-3.546***	-6.484**	-9.241	0.875	-10.994
	(3.201)	(1.304)	(3.140)	(10.126)	(1.889)	(12.213)
1(Year>1869) x (ln(Year)-ln(69)) x 1(Fiction)	16.016***	7.452***	3.178	19.374*	1.962	26.407***
	(4.216)	(2.306)	(3.110)	(10.433)	(3.744)	(8.089)
1(Year>1879) x (ln(Year)-ln(79)) x 1(Fiction)	-6.316	-5.775**	1.642	-2.890	-0.514	-8.229
	(5.585)	(2.705)	(3.228)	(7.088)	(4.805)	(5.784)
1(Year>1889) x (ln(Year)-ln(89)) x 1(Fiction)	-0.093	-0.280	1.032	-8.576	-2.867	3.145
	(5.285)	(2.212)	(3.423)	(5.535)	(3.718)	(5.486)
Constant	1.818	-1.051***	2.112*	-6.482	-3.565***	-1.284
	(1.503)	(0.385)	(1.257)	(3.961)	(0.918)	(1.767)
Observations	160	160	160	160	160	160
R-squared	0.892	0.907	0.891	0.866	0.897	0.959

Notes: The table reports regressions of the natural logarithm of the relative annual frequency (per million words + 0.01) of use of a given word or phrase on the natural logarithm of a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, interactions of these indicators with the difference between the natural logarithm of the current year and the natural logarithm of 29, 39, 49, 59, 69, 79 and 89, respectively, and interactions of all these previous terms with an indicator for whether an observation pertains to fiction books as opposed to non-fiction books. There are two observations per year, one based on the corpus of non-fiction books, and the other on the corpus of non-fiction books (N=160). Each regression is limited to one word or phrase as indicated in the corresponding columns. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.10.